Contraintes préférentielles et ordre des mots en français

Juliette Thuilier

Université Paris Diderot (Paris 7)
École doctorale de Sciences du Langage n°132
U.F.R. Linguistique

Numéro attribué par la bibliothèque

Thèse

Nouveau régime

Pour obtenir le grade de
Docteur en Sciences du Langage
Discipline : Linguistique Générale

Présentée et soutenue publiquement
par

Juliette Thuilier

Le 28 septembre 2012

Contraintes préférentielles et ordre des mots en français

Directeurs de thèse :
Laurence Danlos et Benoît Crabbé

Composition du jury :

Anne Abeillé	Université Paris Diderot
Philippe Blache	CNRS - LPL (rapporteur)
Benoît Crabbé	Université Paris Diderot
Laurence Danlos	Université Paris Diderot
Pollet Samvelian	Université Paris Sorbonne Nouvelle
Shravan Vasishth	Universität Potsdam (rapporteur)

Je tiens à exprimer ma reconnaissance à Laurence Danlos, qui a accepté de diriger mes recherches sans me connaître et m'a accueilli à Alpage les bras ouverts. Sa rigueur scientifique, ses remarques pertinentes et son soutien m'ont beaucoup aidée pour la réalisation de cette thèse. Je tiens ensuite à exprimer mes remerciements à Benoît Crabbé, qui a co-encadré mon travail au cours de ces quatre années. Ce fut un grand plaisir de travailler avec lui pendant cette période. Le travail présenté ici a nettement été influencé par nos discussions et nos collaborations. Je tiens à le remercier également pour l'enthousiasme scientifique et la disponibilité personnelle dont il a fait preuve.

Je remercie Shravan Vasishth et Philippe Blache d'avoir accepté d'être les rapporteurs de cette thèse, ainsi qu'Anne Abeillé et Pollet Samvelian d'avoir accepté d'être membres du jury. Mon travail a beaucoup bénéficié des remarques et des apports scientifiques d'Anne Abeillé. Elle a également procédé à une relecture attentive de ma travail qui a permis d'en améliorer la qualité. C'est grâce à Pollet Samvelian et aux lectures qu'elle m'a suggérées pendant mon Master, que je me suis intéressée aux problématiques abordées dans ce travail. Elle a beaucoup influencé mon cheminement linguistique. Je lui suis aussi très reconnaissante de m'avoir dirigée vers Paris 7 et Alpage, à la fin de mon Master.

Je tiens à exprimer ma reconnaissance à Jean-Marie Marandin pour les discussions riches et stimulantes que nous avons eues, ainsi que pour la relecture attentive de mon travail de thèse. Son enthousiasme et sa disponibilité doivent être salués ici. Un très grand merci aussi à Gwendoline Fox, avec qui j'ai eu la chance de collaborer au début de mes recherches. Nos échanges et notre collaboration ont été très fructueuses pour moi et le travail sur les adjectifs lui doit beaucoup.

Je veux remercier Delphine Tribout pour la validation du classement morphologique des adjectifs et Sarra El Ayari pour avoir mis en place, dans un temps record, un questionnaire en ligne. Merci à toutes les deux pour leur soutien et leur gentillesse. Je remercie les membres d'Alpage, en particulier Djamé Seddah pour le temps passé à m'écouter et à me conseiller autour d'une cigarette pendant la rédaction. Merci aussi à Valérie Hanoka pour son enthousiasme, sa gentillesse et son aide logistique dans les derniers moments. Je remercie également les doctorants et autres membres LLF qui ont, de près ou de loin, participé à ma vie scientifique et professionnelle. Merci également aux quatre étudiants stagiaires qui ont annoté certains des corpus nécessaires à la réalisation de cette thèse : Mathilde, Audrey, Kevin et Fabien.

Merci à Laurent Demaret qui a pris la peine de relire le chapitre concernant les méthodes statistiques et m'a apporté des commentaires très constructifs. Je remercie Manuela pour sa gentillesse, sa disponibilité et, bien sûr, pour son travail de relecture de grande qualité. Cette thèse a beaucoup gagné en qualité d'écriture grâce à elle. Merci à Jean-Marie pour sa relecture attentive.

Enfin, merci à Mathieu pour son soutien, sa patience et surtout sa présence. J'espère être à mon tour à la hauteur pour l'accompagner et le soutenir pendant la rédaction de sa thèse.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'objectif de cette thèse est d'étudier la notion de contrainte préférentielle en syntaxe, à travers deux problèmes d'ordre des mots du français. Les contraintes préférentielles peuvent se définir comme des contraintes agissant sur l'acceptabilité des phrases d'une langue et non sur leur grammaticalité.

Dans la tradition de la syntaxe générative, ce type de contraintes n'a pas été étudié, car ces dernières ont été envisagées comme des contraintes de performance et non de compétence. En effet, l'objectif initial de la syntaxe générative a été de modéliser la compétence des locuteurs d'une langue en définissant un système de règles qui génère l'ensemble des séquences bien-formées et qui rejette celles qui sont mal-formées. C'est donc l'étude de l'opposition entre grammaticalité et agrammaticalité qui a orienté les études menées en syntaxe, fondé la méthodologie de recherche et entraîné l'élaboration de méthodes formelles permettant de rendre compte des phénomènes étudiés. Ces recherches ont permis des avancées considérables dans la description, la compréhension et la modélisation du système des langues. Cependant, il semble qu'en se cantonnant au contraste grammatical/agrammatical, le domaine de la syntaxe n'embrasse pas la totalité des phénomènes qui expliquent la manière dont les mots et les constituants sont agencés. Il ne rend donc pas compte de certaines propriétés de la langue. L'étude des contraintes préférentielles a été développée dans les travaux de Wasow (1997, 2002) sur l'ordre des mots, puis poursuivie dans un certain nombre de travaux traitant de l'alternance dative en anglais (Bresnan *et al.*, 2007; Bresnan & Ford, 2010; Bresnan & Nikitina, 2009; Gries, 2003a, parmi d'autres). Ce phénomène est présenté dans l'exemple (1).

(1) Alternance dative
   a. Construction à SP datif : *John gave toys to the children*
   b. Construction à double objet : *John gave the children toys*

Afin d'illustrer l'idée de préférence en syntaxe, nous prenons les exemples (2-a) et (2-b) qui sont deux phrases grammaticales de l'anglais. Si l'on compare l'acceptabilité de ces deux phrases, il apparaît que la phrase (2-a) est plus naturelle que la phrase (2-b).

(2) a. *He gave me the backpack*
b. *He gave the backpack to me*

Du point de vue de la syntaxe générative, les deux séquences sont équivalentes, puisque grammaticales, et doivent être acceptées par la grammaire. Or, si la préférence pour l'ordre présenté dans la première phrase est écartée de l'objet de la syntaxe, la description des propriétés spécifiques à la langue ne semble pas complète.

De plus, la distinction nette entre contrainte de compétence, touchant à la grammaticalité des phrases, et contrainte de performance, n'affectant que leur acceptabilité, n'est pas toujours facile à établir. Dans le cas de l'alternance dative, certaines séquences sont jugées agrammaticales, alors qu'elles peuvent être acceptables dans des contextes spécifiques. Par exemple, les phrases (3-a) et (3-b) sont reportées avec les jugements qui leur sont généralement associés et selon lesquels des verbes exprimant une façon de parler, tels que *mutter* ('marmonner'), ne peuvent pas être suivis de la construction à double objet. Cependant, la séquence considérée comme agrammaticale apparaît dans l'exemple (3-c), tiré de Bresnan & Nikitina (2009), qui semble être une phrase acceptable de l'anglais.

(3) a. *Susan muttered the news to Rachel*
b. **Susan muttered Rachel the news*
c. *Shooting the Urasian a surprised look, she* ***muttered him a hurried apology*** *as well before skirting down the hall.*

Il semble donc que la limite entre séquences grammaticale et agrammaticale soit relativement floue. Prendre en compte la dimension de l'acceptabilité et par conséquent des contraintes préférentielles permet de surmonter les problèmes soulevés par la classification binaire des phrases. Les travaux précédemment cités ont permis de mettre à jour, pour l'alternance dative, le rôle de facteurs hétérogènes, tels que la longueur relative des compléments, leur statut discursif, leur caractère pronominal ou animé... Nous désignons sous le terme de contrainte préférentielle ce type de facteur.

Ce travail se concentre sur l'étude des contraintes préférentielles. On admettra que ces dernières constituent des propriétés spécifiques à la langue. Nous supposons donc que l'étude d'une langue passe par la description, la compréhension et la formalisation des contraintes préférentielles.

L'étude des contraintes préférentielles pose néanmoins un problème méthodologique central : comment les observer et en rendre compte ? Les outils traditionnels de la linguistique ne sont pas adaptés dans la mesure où ils ont été conçus dans la perspective d'une opposition binaire. Il faut donc envisager de nouveaux outils et de nouvelles méthodes. L'étude des contraintes préférentielles peut s'envisager à partir de deux types de données : les données extraites de corpus, à partir desquelles on observe des tendances, et les données recueillies à l'aide de protocoles expérimentaux. L'émergence croissante de corpus richement annotés permet de recueillir un nombre important de données relatives à un problème syntaxique spécifique. Leur maniement et leur traitement exigent également des moyens technologiques et computationnels

qui sont aujourd'hui facilement accessibles.

Sur le plan méthodologique, nous nous situons dans la lignée des travaux conduits sur l'alternance dative en anglais et en particulier ceux de Bresnan *et al.* (2007) et Bresnan & Ford (2010). Ces derniers ont mis en avant l'utilisation de l'analyse de données de corpus grâce à la statistique inférentielle. Ces travaux fondateurs pour la problématique qui nous occupe, montrent que l'on peut espérer établir des généralités sur des questions de préférence à partir des méthodes d'analyse de données. Ils soulignent également la complémentarité des données de corpus et du travail expérimental, en mettant en lumière la corrélation entre les préférences dégagées sur corpus et les préférences observées chez les locuteurs.

Phénomènes d'ordre en français

À notre connaissance, les travaux présentés dans cette thèse sont les premiers à exposer une étude sur le français dans cette perspective et avec cette méthodologie. À la différence des travaux de Bresnan *et al.*, nous nous intéressons à des phénomènes d'ordre des mots et non d'alternance de constructions. Plus précisément, la notion de contrainte préférentielle sera mise en oeuvre à travers l'étude de deux phénomènes : la position de l'adjectif épithète par rapport au nom et l'ordre relatif des compléments sous-catégorisés par le verbe dans le domaine postverbal.

Le premier a fait l'objet de nombreux travaux et constitue un phénomène très bien documenté. Les différents facteurs influençant la position de l'adjectif ont été largement discutés. Il est alors apparu naturel d'entreprendre une étude sur des données attestées ayant pour objectif la modélisation de l'effet et de l'interaction des différentes contraintes dégagées dans la littérature. Afin de poursuivre la problématique liée à l'ordre des constituants en français, nous nous sommes intéressée à l'ordre des compléments postverbaux. Ce phénomène, peu étudié en français, présentait l'avantage de pouvoir être mis en parallèle avec des phénomènes comparables dans d'autres langues, telles que l'anglais ou l'allemand.

La position de l'adjectif épithète

La position de l'adjectif épithète par rapport au nom n'est pas fixe en français, comme le montre l'exemple (4).

(4) a. *Voilà vous savez tout, ha non vous ne savez pas tout, un* ***délicieux bourgogne*** *accompagnait ce petit repas*[1]

b. *Tartines gourmandes avec un* ***Bourgogne délicieux*** *recommandé par Marie à la librairie Cave à vins "Les Flo des mots" à Sète*[2]

1. http://boubouatable.over-blog.com/article-le-temps-des-cerises-86243760.html, page consultée le 6 juin 2012.

2. http://biblavardac.blogspot.fr/2009/10/villages-et-cites-du-livre-en-france.html, page consultée le 6 juin 2012.

Dans ce phénomène, il existe une seule contrainte qui impose une position et agit donc réellement sur la grammaticalité du syntagme nominal (SN) : la présence d'un dépendant postadjectival dans le syntagme adjectival (SAdj). Lorsqu'un adjectif est accompagné d'un dépendant, il est obligatoirement postposé au nom, comme le montre l'opposition entre (5-a) et (5-b).

(5) a. *une mère fière de son fils*
b. **une fière de son fils mère*

Cette règle ne souffre aucune exception. En revanche, l'ensemble des autres contraintes influant sur la place de l'adjectif agit de façon préférentielle et non catégorique. Par exemple, une partie des adjectifs appartenant à la classe sémantique des intensionnels, c'est-à-dire des adjectifs qui modifient la relation entre le nom et le référent qu'il désigne, ont une préférence lexicale pour la position antéposée. C'est le cas de l'adjectif *prétendu*, en (6-a), qui instaure une distance entre le référent du SN et le nom *complot*. De même, l'adjectif *futur*, en (6-b), signale que l'étiquette *président* n'est pas adaptée au référent désigné au moment où le SN est produit.

(6) a. *le prétendu complot*
b. *le futur président*

Cependant, les adjectifs appartenant à cette classe n'apparaissent pas obligatoirement dans cette position, comme en attestent les exemples (7-b) et (7-a). Cela signifie que la qualité d'adjectifs intensionnels n'impose pas une position, mais la favorise très fortement.

(7) a. *Voilà le* ***complot prétendu*** *et l'idéologie de classe que les auteurs mettent en images.*[3]
b. *Il a pris soin au préalable de "rappeler que le* ***président futur*** *de l'UMP qui sera élu à l'automne, ce n'est pas celui qui sera le candidat assuré de 2017"*[4]

Par ailleurs, il existe des contraintes qui assouplissent les préférences lexicales. C'est le cas par exemple de la coordination d'adjectifs. En effet, un adjectif ayant une forte préférence pour l'antéposition peut apparaître beaucoup plus facilement en postposition, lorsqu'il est coordonné. Nous reprenons ici l'exemple proposé par Abeillé & Godard (1999), avec les adjectifs intensionnels *vrai* et *faux*.

(8) a. *des faux coupables / ?des coupables faux*
b. *des vrais coupables / ?des coupables vrais*
c. *des coupables vrais ou faux*

3. http://quilukru.blogspot.fr/2012/02/les-nouveaux-chiens-de-garde-critique.html, page consultée le 7 juin 2012.

4. http://www.lemonde.fr/election-presidentielle-2012/article/2012/05/16/a-la-presidence-de-l-ump-jean-francois-cope-veut-un-homme-comme-lui_1702349_1471069.html, page consultée le 7 juin 2012.

Les séquences où ces adjectifs sont postposés seuls au nom *coupable* sont difficilement acceptables, au moins hors contexte. En revanche, une fois coordonnés, les deux adjectifs peuvent être postposés au nom de façon très naturelle.

La position de l'adjectif ne répond donc pas à des contraintes catégoriques imposant une position spécifique, mais à des contraintes qui favorisent une position plutôt que l'autre. Notre objectif sera de montrer que l'approche en termes de contraintes préférentielles est adaptée pour ce phénomène dans la mesure où elle permet de rendre compte de son caractère multidimensionnel.

Ordonnancement des compléments postverbaux

L'ordre des constituants postverbaux est relativement libre en français. Le verbe constitue le point de référence autour duquel les constituants s'organisent linéairement. Les éléments qui succèdent au verbe ne se voient pas imposer d'ordre en termes de la grammaticalité de la séquence. Cette liberté d'ordonnancement connaît deux limites. Premièrement, les noms nus compléments ne peuvent pas être séparés du verbe par un autre constituant, comme le montrent les exemples (9), extraits de Abeillé & Godard (2006, p. 13).

(9) a. **Ce lieu fait aux enfants* ***peur***
b. **Le président rendra aux victimes* ***hommage***

Deuxièmement, le français dispose d'adverbes "légers", selon les termes de Abeillé & Godard (2001), qui apparaissent obligatoirement adjacents au verbe, comme en atteste l'agrammaticalité des exemples (10) (tirés de Abeillé & Godard, 2001).

(10) a. **Paul va au cinéma* ***trop***
b. **Marie comprend le cours* ***bien***

Hormis ces deux exceptions, les autres constituants postverbaux ne sont pas ordonnés selon des contraintes catégoriques définissant la grammaticalité de la séquence. Nous émettons l'hypothèse selon laquelle l'ordre choisi par les locuteurs est influencé par différentes contraintes préférentielles. Nous nous concentrons plus spécifiquement autour des contraintes intervenant dans l'ordonnancement des constituants sous-catégorisés par la tête verbale. Le travail que nous proposons aura pour objectif premier d'identifier les contraintes générales qui ont une influence sur ce phénomène.

Plan de la thèse

Cette thèse est composée de six chapitres. Les deux premiers présentent de façon théorique la notion de contraintes préférentielles et les méthodes déployées pour leur étude. Les quatre derniers sont regroupés en deux grandes parties : une partie consacrée au problème de la position de l'adjectif en français (Chapitres 3 et 4) et une autre dédiée à l'ordre des compléments postverbaux (Chapitres 5 et 6).

Chapitre 1 L'idée que la syntaxe doit rendre compte de contraintes préférentielles ne va pas de soi dans la mesure où ces dernières n'affectent pas la grammaticalité des phrases. Nous justifions l'intégration de cette dimension au domaine traitant de la compétence langagière, en nous appuyant sur l'*argument typologique* développé par Bresnan (2007a). D'après cet argument, certaines contraintes s'expriment de façon catégorique dans certaines langues et de façon préférentielle dans d'autres. Cela signifie que la même contrainte affecte l'acceptabilité, et donc la performance, dans certaines langues, tandis qu'elle touche à la grammaticalité, et donc à la compétence, dans d'autres. Nous supposons alors que la nature des contraintes (compétence ou performance) ne diffère pas d'une langue à l'autre et qu'un seul type de contrainte s'exprime de façon catégorique dans certaines langues et de façon préférentielle dans d'autres. De plus, les travaux de Bresnan & Hay (2008) et Bresnan & Ford (2010) montrent que les locuteurs de différentes variétés de l'anglais (américain, australien et néo-zélandais) ne présentent pas tout à fait les mêmes préférences dans le phénomène de l'alternance dative. À moins de postuler que les phénomènes de performance sont différents pour ces trois variétés d'anglais, il faut intégrer les préférences dans les connaissances langagières des locuteurs afin de rendre compte des variations relatives à ce phénomène.

Nous émettons l'hypothèse que les contraintes préférentielles interviennent dans le cas où aucune règle catégorique ne s'applique, c'est-à-dire lorsqu'il existe un espace de liberté en termes de grammaticalité. Leur étude constitue donc un aspect complémentaire de la description et de la formalisation du système des langues conçu autour de l'opposition grammatical/agrammatical. D'un point de vue méthodologique, les trois outils disponibles pour l'étude des phénomènes syntaxiques, à savoir les jugements de grammaticalité, l'analyse des données de corpus et les expériences psycholinguistiques, constituent des ressources complémentaires, dans la mesure où chacun d'entre eux permet de dépasser les limites des deux autres. L'étude des contraintes préférentielles passe par l'utilisation des deux derniers outils, comme l'illustrent les travaux sur l'alternance dative en anglais (Bresnan *et al.*, 2007; Bresnan & Ford, 2010; Bresnan & Hay, 2008; Bresnan & Nikitina, 2009; Collins, 1995; Gries, 2003b; Snyder, 2003; Thompson, 1990). Une partie de ces travaux a ouvert la voie à l'élargissement de l'objet de la syntaxe à des dimensions non-catégoriques, en mettant en oeuvre une machinerie statistique permettant d'envisager la possibilité de passer d'observations de tendances en corpus à des propriétés du système de la langue.

Chapitre 2 Une fois définie la notion de contrainte préférentielle, émerge une question centrale, celle de la méthodologie. Étant donné que les contraintes préférentielles échappent aux méthodes traditionnelles de la syntaxe, il est nécessaire de définir en détail les outils permettant leur description et leur analyse. Le chapitre 2 se concentre principalement sur l'obtention et les méthodes d'analyse des données de corpus. L'utilisation de ces données pose deux problèmes majeurs : la représentativité et la généralisation. Premièrement, un corpus, c'est-à-dire une collection de phrases produites dans des conditions non-expérimentales, est un échantillon de la langue.

Or, pour pouvoir observer des propriétés générales de la langue dans l'échantillon, il est nécessaire que ce dernier soit représentatif de la langue, de la même façon que l'échantillon des personnes interrogées dans un sondage d'opinion doit être représentatif de la population suivant un certain nombre de critères (sexe, âge, catégories socio-professionnelles…). Les études de cas proposées n'échappent pas au problème de la représentativité, mais l'utilisation de corpus variés et de données expérimentales a pour but de limiter au maximum les biais dus à l'échantillonnage des données de corpus. Deuxièmement, la généralisation pose le problème du passage des observations sur corpus à des propriétés de la langue. Il est notamment nécessaire de s'assurer que les propriétés observées ne sont pas des artéfacts liés à l'échantillonnage. Pour cela, nous utilisons des méthodes statistiques inférentielles qui sont largement utilisées dans d'autres domaines de la linguistique et dans d'autres sciences humaines. Nous employons notamment la régression logistique qui permet de modéliser la probabilité qu'une construction ou un ordre soit réalisé, étant donné un ensemble de contraintes linguistiques intégrées au modèle sous la forme de variables prédictrices. Nous utilisons également les modèles à effets mixtes, qui permettent de prendre en compte les variables formant des groupes dans les données. Dans le cas de la modélisation des phénomènes linguistiques, cela permet notamment de tenir compte des spécificités relatives aux items du lexique. Au-delà de l'explicitation des aspects techniques, nous cherchons à montrer l'intérêt de ces méthodes pour l'étude du système de la langue et plus précisément des contraintes préférentielles en syntaxe. Enfin, les données issues de protocoles expérimentaux sont complémentaires des données de corpus. Elles permettent notamment de pallier les limites de représentativité. Deux types d'expérimentation sont évoqués, en lien avec leur utilité pour dépasser les limites intrinsèques aux données de corpus.

Chapitre 3 Ce chapitre est dédié au bilan des contraintes proposées dans la littérature pour expliquer l'alternance de position de l'adjectif épithète. L'hypothèse émise est qu'il s'agit d'un phénomène multifactoriel et que la quasi-totalité de ces facteurs représente des contraintes préférentielles.

Dans ce chapitre, sont passés en revue les aspects phonologiques, lexicaux, syntaxiques, sémantiques et discursifs. À partir d'un ensemble d'exemples attestés, nous chercherons à montrer que la position de l'adjectif n'agit pas sur la grammaticalité des phrases, mais plutôt sur leur acceptabilité. Seule la présence d'un dépendant postadjectival, comme en (5-b), est identifiée comme une contrainte catégorique affectant la grammaticalité de la phrase. La nature préférentielle des autres contraintes justifie l'intérêt d'envisager ce problème à l'aune des méthodes décrites dans le chapitre 2.

Chapitre 4 Ce chapitre est consacré à l'étude sur corpus de l'alternance de position de l'adjectif dans le syntagme nominal. À partir de données extraites automatiquement du *French Treebank* (Abeillé & Barrier, 2004), nous avons constitué une table de données dans laquelle une grande partie des facteurs relevés dans le chapitre 4 ont été captés à l'aide de variables annotées. Nous proposons une analyse de ces données

organisée sous la forme de comparaison de modèles permettant de mettre en lumière l'importance des caractéristiques lexicales de l'adjectif, mais aussi de l'identité du nom avec lequel il se combine. Les facteurs relatifs à la configuration du syntagme adjectival et du syntagme nominal sont également significatifs lorsqu'ils sont combinés aux contraintes relevant du lexical. Cette approche novatrice permet de prendre en compte le caractère multifactoriel du phénomène et d'en proposer une formalisation sous la forme d'un modèle statistique qui prédit l'ordre attesté pour 87% des adjectifs alternant dans nos données. Grâce à un questionnaire permettant l'élicitation des préférences de locuteurs, nous montrons que, pour 29 phrases, les probabilités d'antéposition estimées en corpus sont corrélées aux proportions de sujets préférant la position antéposée, ce qui laisse supposer que les observations faites en corpus sont en correspondance avec une forme de connaissance langagière.

Chapitre 5 Ce chapitre concerne l'ordre des compléments postverbaux en français. Étant donné que peu de travaux se sont concentrés sur les contraintes entrant en jeu dans ce phénomène pour le français, nous proposons un bilan des facteurs identifiés dans l'ordonnancement des dépendants du verbe d'autres langues. Il existe de grandes tendances observées pour un nombre important de langues et selon lesquelles les référents animés précèdent les non-animés, les constituants pronominaux précèdent les non-pronominaux et les éléments définis précèdent les indéfinis. De plus, pour les langues Verbe - Objet comme le français, les constituants courts et grammaticalement simples tendent à précéder les constituants longs et complexes. La présentation des facteurs est organisée autour de trois niveaux : le poids grammatical, les aspects lexico-sémantiques et les aspects discursifs. Les différentes contraintes sont présentées avec l'objectif de rendre compte de la manière dont ces dernières agissent sur l'ordre des dépendants du verbe.

En ce qui concerne le français, il ressort des travaux exposés qu'il n'existe pas de contrainte catégorique guidant l'ordre des mots, hormis le cas des noms nus (exemple (9), Abeillé & Godard, 2006). Certaines contraintes ont été identifiées comme ayant une influence sur l'ordre attesté : la longueur et la complexité syntaxique (Abeillé & Godard, 2006; Berrendonner, 1987; Blinkenberg, 1928) et la sémantique du verbe (Schmitt, 1987a,b). Concernant la structure informationnelle, il n'existe pas d'étude permettant d'évaluer l'influence de facteurs tels que le caractère défini des constituants ou l'opposition *information donnée* - *information nouvelle* (Prince, 1981), alors que l'impact de ces éléments a été mis à jour pour des phénomènes d'ordre en anglais et en allemand. Cet état de l'art fait apparaître que l'ordonnancement des compléments postverbaux du français est un problème qui doit être exploité à l'aide de données de corpus et d'un travail quantitatif permettant de mettre à jour les contraintes préférentielles guidant le choix pour un ordre ou pour l'autre.

Chapitre 6 L'étude proposée dans ce chapitre repose sur des données extraites de corpus écrits et oraux ainsi que sur deux questionnaires d'élicitation de jugements de grammaticalité. À partir de l'examen d'une première table de données, nous propo-

sons une discussion de la notion de poids grammatical en français, en nous inspirant du travail de Wasow (1997, 2002). Les données révèlent que la mesure en nombre de mots est une bonne estimation du poids des constituants. Après avoir montré que l'item verbal a une influence sur l'ordonnancement des compléments, une deuxième table de données mieux échantillonnée est proposée. Dans cette table, sont annotés le caractère animé des référents selon les catégories de Zaenen *et al.* (2004) et les classes sémantiques des verbes d'après le dictionnaire *Les Verbes du Français* (Dubois & Dubois-Charlier, 1997). La modélisation statistique des données de cette deuxième table montre que le phénomène étudié est significativement influencé par la longueur relative des compléments ainsi que par le verbe associé à sa classe sémantique. Nous proposons deux pistes d'analyse permettant de comprendre la manière dont se façonnent les préférences verbales en faveur d'un ordre ou de l'autre.

D'après le modèle statistique construit, nous ne disposons pas d'éléments prouvant l'effet du caractère animé et du caractère défini sur l'ordre des compléments, ce qui semble distinguer le français d'autres langues telles que l'anglais ou l'allemand. Premièrement, l'analyse des données de corpus ne présente aucune preuve permettant d'affirmer que les contraintes relatives au caractère animé des référents ont un impact sur l'ordre des compléments. Ce constat semble être conforté par la non-significativité de l'interaction du caractère animé et de l'ordre des compléments sur les jugements des locuteurs recueillis par l'intermédiaire d'un questionnaire. Deuxièmement, aucun effet du caractère défini des compléments postverbaux n'émerge dans l'analyse des données. Enfin, la distinction entre référent donné ou nouveau (Prince, 1981), annotée sur une sous-partie de la table de données, ne présente pas d'effet significatif sur l'ordre des compléments du verbe. De plus, dans le même ordre d'idées que pour le caractère animé, il n'y a pas d'effet observable de l'interaction entre l'ordre et le caractère donné ou nouveau des référents sur les jugements de locuteurs, également recueillis à l'aide d'un questionnaire.

Cette étude exploratoire met en lumière l'influence du poids grammatical et des verbes désambiguïsés à l'aide des classes sémantiques en contexte. La formalisation du phénomène est proposée sous la forme d'un modèle statistique.

Chapitre 1

Les contraintes préférentielles

Sommaire

Dans cette thèse, nous nous intéressons à la notion de préférence et nous étudions les contraintes préférentielles qui interviennent dans l'ordonnancement des mots et des constituants en français. Nous émettons l'hypothèse que les préférences sont des propriétés de la langue. Par conséquent, dans le cadre de la description et de la formalisation des langues, il faut étudier ces préférences.

La prise en compte des contraintes préférentielles dans le champ de la syntaxe pose deux problèmes principaux : un problème théorique et un problème méthodologique. Premièrement, considérer que les préférences font partie de la connaissance du langage ne va pas de soi, car les contraintes préférentielles affectent l'acceptabilité des phrases et non leur grammaticalité. Elles ne font donc pas partie du champ de recherche de la syntaxe, tel qu'il est défini par exemple par Chomsky (1965). Il est nécessaire de fournir des arguments pour justifier l'élargissement de l'objet d'étude de la syntaxe en intégrant ces contraintes. Deuxièmement, les contraintes préférentielles ne peuvent pas être mises à jour au moyen des méthodes traditionnelles utilisées en syntaxe (introspection, jugement de grammaticalité). Leur étude implique l'utilisation de corpus annotés et d'expériences psycholinguistiques, méthodes qui posent des problèmes de représentativité et de généralisation.

Ce chapitre s'organise en quatre sections. Dans la première, nous tracerons les contours de la notion de contrainte préférentielle en explicitant ses principales caractéristiques et en montrant l'intérêt de l'étude de cette notion. La deuxième section sera consacrée au problème de la méthode et de la généralisation dans l'étude des contraintes préférentielles. Dans la troisième partie, nous présenterons plus en détail les différentes études menées sur l'alternance dative en anglais, dans la mesure où ces études ont mis en oeuvre la méthodologie et les notions qui nous intéressent. Enfin, dans la dernière section, nous établirons un bilan concernant l'objet de la syntaxe, une fois prises en compte les contraintes préférentielles.

1.1. Qu'est-ce qu'une contrainte préférentielle ?

Nous introduisons le problème des préférences en syntaxe à l'aide d'un exemple : la position de la particule dans les constructions verbe-particule en anglais. Dans ce type de construction, la particule peut être directement adjacente au verbe (1-a) ou apparaître après le SN objet (1-b).

(1) a. *We **figure out** the problem*
b. *We **figure** the problem **out***

L'ordre de la particule et du SN complément est relativement libre et le choix d'un ordre ou de l'autre dépend notamment de la longueur et de la complexité du SN (Wasow, 1997) et de la relation sémantique entre le verbe et la particule (Wasow & Arnold, 2003).

Cependant, ce type de phénomène ne fait pas partie du champ de recherche du grammairien, tel qu'il est défini par Chomsky (1965). En effet, la démarche généra-

tive s'appuie sur l'opposition grammatical/agrammatical pour « *déterminer, à partir des données de la performance, le système sous-jacent de règles qui a été maîtrisé par le locuteur-auditeur et qu'il met en usage dans sa performance effective* » [1]. Or, l'ordre des éléments dans la construction verbe-particule n'affecte pas la grammaticalité, mais l'acceptabilité de la phrase, comme l'explique Chomsky (1965) à partir de l'exemple (2).

(2) *I **called** the man who wrote the book that you told me about **up***

Selon Chomsky, il s'agit d'un exemple de phrase grammaticale mais inacceptable, pour laquelle les causes d'inacceptabilité ne relèvent pas de la grammaire de la langue, mais de facteurs extra-grammaticaux tels que les limitations de la mémoire, les facteurs stylistiques ou discursifs.

Nous émettons ici l'hypothèse que l'étude des préférences en syntaxe constitue une démarche complémentaire et compatible avec la démarche générativiste. Aux contraintes catégoriques qui définissent des structures du langage, s'ajoutent des contraintes préférentielles qui sont à l'oeuvre dans la production et la compréhension langagière. En s'intéressant aux préférences, on cherche à mieux décrire, formaliser et comprendre les mécanismes et les règles mises en oeuvre par les locuteurs-auditeurs d'une langue. Étant donné que les préférences n'affectent pas la grammaticalité, elles ne peuvent pas être capturées par les méthodes traditionnelles de la syntaxe. Leur étude implique notamment l'utilisation de corpus richement annotés. La necessité de tels corpus explique, en partie, que l'intérêt pour les contraintes préférentielles en syntaxe émerge avec le renouveau "technologique" des années 90 : la création de grands corpus annotés et l'utilisation de méthodes statistiques visant à tirer des généralités à partir de l'échantillon que constituent ces corpus. Ces méthodes permettent d'envisager de nouveaux problèmes, jusque là impossibles à étudier, à savoir les phénomènes où plusieurs contraintes interagissent, et agissent donc de façon préférentielle.

1.1.1. Caractéristiques des contraintes préférentielles

Lorsqu'une variable intervient de façon non-catégorique dans un phénomène linguistique, on parle de contrainte préférentielle. Par exemple, dans le cas d'une alternance entre deux constructions, A et B, une variable binaire agit de façon préférentielle si l'une de ses deux valeurs potentielles se rencontre plus fréquemment avec la construction A qu'avec la construction B, et ce de façon statistiquement significative. Ce type de contraintes n'est pas intrinsèquement préférentiel : une contrainte peut être catégorique dans une langue et préférentielle dans une autre langue, comme nous le montrerons à partir d'exemples dans la section 1.1.2.2. Ces contraintes peuvent correspondre à des variables continues (longueur d'un constituant, fréquence d'un

1. *« …to determine from the data of performance the underlying system of rules that has been mastered by the speaker-hearer and that he puts to use in actual performance.* » (Chomsky, 1965, p. 4)

lemme) ou à des variables discrètes (caractère animé, personne grammaticale). Elles peuvent se situer à différents niveaux : prosodie, phonologie, syntaxe, sémantique, structure informationnelle. Enfin, elles interviennent dans des phénomènes où il y a un choix entre plusieurs réalisations : alternance de constructions, ordre relatif de constituants, choix entre plusieurs formes d'un mot…

Nous nous intéressons aux contraintes préférentielles intervenant dans l'ordonnancement des mots et des constituants. Nous nous plaçons dans la lignée des travaux qui séparent règles de constituance et règles de précédence linéaire (Falk, 1983; Gazdar & Pullum, 1981). Nous considérons donc que les règles de formation des constituants sont indépendantes des règles de linéarisation des constituants. Les contraintes préférentielles agissent au niveau de la précédence linéaire, quand aucune règle catégorique ne s'applique et que l'ordre relatif des éléments est considéré comme libre. Nous émettons l'hypothèse qu'une partie des règles de linéarisation est constituée de contraintes préférentielles. Ainsi, l'approche proposée est compatible avec les théories séparant constituance et linéarisation, comme par exemple *Head-driven phrase structure* (Pollard & Sag, 1994) et les travaux qui ont suivi : Reape (1994, 1996), Donohue & Sag (1999) et Kathol (2000), parmi d'autres.

Ces contraintes sont censées intervenir à la fois en production et en compréhension, comme en atteste une série de travaux sur l'alternance dative en anglais. Cette alternance, exemplifiée en (3), se caractérise par la possibilité d'un choix entre une construction à SP datif et une construction à double objet.

(3) a. construction à SP datif : *He gives [the picture] [to Mary]*
b. construction à double objet : *He gives [Mary] [the picture]*

Nous présentons d'abord deux travaux concernant le phénomène en production. Ensuite, nous en exposons deux autres qui se centrent sur la compréhension.

Premièrement, Bresnan *et al.* (2007) proposent une analyse de l'alternance dative en s'appuyant sur des données de corpus oraux et écrits, dont ils font émerger des contraintes préférentielles, à partir de méthodes d'analyse sur lesquelles nous reviendrons plus particulièrement dans le chapitre 2. Ce travail montre que les contraintes préférentielles sont observables dans des textes produits dans des conditions non-expérimentales, ce qui laisse supposer qu'elles interviennent naturellement dans la production. Le travail de Tily *et al.* (2009), fait à partir d'un corpus d'oral, montre que, dans la parole spontanée, la durée acoustique peut être influencée par les contraintes préférentielles. Plus précisément, lorsqu'il y a production d'une construction à SP datif alors que les contraintes préférentielles favoriseraient une construction à double objet, la durée acoustique de la préposition *to* est plus longue que dans le cas où cette préposition apparaît dans un contexte favorisant la construction à SP datif. Il semble donc que les contraintes préférentielles soient identifiables en production dans des corpus écrits et oraux.

Deuxièmement, Bresnan & Ford (2010) développent une *étude corrélationnelle*[2] à

2. Le terme *étude corrélationnelle* renvoit à une tâche psycholinguistique reposant sur des données attestées qui présentent des corrélations. Ces études sont généralement utilisées dans le cadre

partir de laquelle elles montrent que les contraintes préférentielles affectent le processus de compréhension de l'alternance dative. L'étude corrélationnelle proposée repose sur une tâche de décision lexicale continue [3] (Ford, 1983). Lors de cette tâche, les sujets lisent une phrase mot par mot, à leur rythme et ils doivent décider si chaque mot lu appartient, ou n'appartient pas, à la langue anglaise. Les auteurs recueillent les temps de réaction à la lecture de la préposition *to* dans la construction à SP datif. L'hypothèse sous-jacente à cette étude corrélationnelle est que plus les facteurs en présence favorisent une construction à SP datif, moins les sujets mettront de temps à identifier le mot *to*, et inversement, moins les contraintes préférentielles en présence favorisent la construction à SP datif, plus les sujets auront besoin de temps pour identifier *to* comme un mot. Cela signifie que, lorsqu'ils ne s'attendent pas à rencontrer la préposition *to*, les sujets mettent plus longtemps à l'identifier comme un mot. Les observations de Bresnan & Ford sont conformes à cette hypothèse. Lors de la lecture, les locuteurs semblent avoir des attentes en correspondance avec les contraintes préférentielles intervenant dans le phénomène.

Tily *et al.* (2008) présentent un travail reposant sur la technique de l'oculométrie (traduction de *eye-tracking*) et visant à observer si, en compréhension, les locuteurs ont des attentes en correspondance avec les préférences dégagées dans le travail sur corpus fait par Bresnan *et al.* (2007). Plus précisément, les auteurs cherchent à montrer que les préférences de chaque verbe, pour une construction ou pour l'autre, se retrouvent sous la forme d'attente des locuteurs pour une construction plutôt que pour l'autre lorsqu'ils entendent un verbe. Lors de l'expérience, le sujet, équipé d'un oculomètre (*eye-tracker*) entend une phrase contenant une construction à double objet ou à SP datif, en même temps qu'il regarde une image contenant les trois acteurs intervenant dans la phrase (le sujet, le destinataire et le thème). Tily *et al.* observent que les participants fixent plus rapidement le premier argument postverbal lorsque le biais verbal correspond à la construction effectivement produite. Ce travail constitue un argument en faveur de l'idée que les contraintes préférentielles interviennent dans la compréhension d'énoncés oraux, sous la forme d'attentes syntaxiques.

L'ensemble de ces travaux tend à montrer que les contraintes préférentielles identifiées en corpus ont un rôle important lorsque les locuteurs produisent des phrases contenant l'alternance dative, et lorsqu'ils en analysent, lors de la compréhension orale ou de la lecture.

La notion de contrainte préférentielle est en lien avec celle de *soft constraint* telle qu'elle a été développée et étudiée par Keller (2000) et Sorace & Keller (2005). Nous proposons ici une comparaison des deux notions, en utilisant le terme *contrainte préférentielle* pour désigner la notion au coeur de cette thèse, et le terme anglais *soft constraint* pour faire référence à la notion de Keller (2000) et Sorace & Keller (2005).

Premièrement, Sorace & Keller posent une distinction fondamentale entre *hard*

de travaux de corpus et visent à trouver des corrélations entre les résultats obtenus par l'analyse de données de corpus et les résultats psycholinguistiques. Nous réservons le terme *expérience* aux protocoles dans lesquels les variables prédictrices sont contrôlées et décorrélées.

3. Traduction de *continuous lexical decision task*.

constraints et *soft constraints*. Les premières donnent lieu à des jugements d'inacceptabilité forts et stables, alors que les secondes déclenchent des jugements plus gradués et instables. Dans le cas des contraintes préférentielles, la définition ne fait pas référence à des niveaux de jugement d'acceptabilité. Ces contraintes agissent bien sur l'acceptabilité, mais, à la différence des *soft constraints*, leur nature préférentielle ne dépend pas de la force des jugements d'inacceptabilité : la violation ou la satisfaction des contraintes préférentielles peut déclencher des jugements forts ou modérés. La qualité préférentielle est définie en fonction de la fréquence avec laquelle une contrainte se rencontre dans un phénomène. En effet, nous considérons qu'à partir du moment où le choix d'une structure ou d'un ordre ne peut pas s'expliquer en termes catégoriques, le phénomène est régi par une ou plusieurs contraintes préférentielles. Les préférences et leur force sont alors liées à la fréquence de satisfaction (ou violation) de chacune des contraintes intervenant dans ce phénomène.

Deuxièmement, Sorace & Keller (2005) postulent que la distinction *hard* vs. *soft constraints* ne connaît pas de variation entre les langues : « *nous prédisons qu'il n'y a pas de contraintes qui soient* soft *dans une langue et* hard *dans une autre* » [4]. Cela va à l'encontre d'une propriété que nous avons énoncée précédemment et qui est définitoire pour la notion que nous étudions : une contrainte n'est pas intrinsèquement préférentielle et sa réalisation peut varier selon les langues [5].

Enfin, Sorace & Keller établissent deux autres propriétés relatives aux *soft constraints* : elles sont dépendantes du contexte linguistique et présentent un comportement particulier dans l'acquisition et l'attrition de la première langue et des langues secondes. Keller (2000, p. 126) définit la notion de contexte comme les phénomènes grammaticaux interphrastiques tels que la structure informationelle et la référence. Dans la mesure où les contraintes préférentielles étudiées ici renvoient en partie au contexte tel qu'il vient d'être défini, on peut considérer que les phénomènes d'ordre que nous étudions sont sensibles au contexte et qu'une partie des contraintes préférentielles définit ce contexte. Par exemple, dans le cas de l'alternance dative, que nous détaillerons dans la section 1.3, le choix de la construction dépend en partie du caractère donné ou nouveau des référents impliqués. En ce qui concerne l'acquisition et l'attrition, nous ne disposons pas d'éléments permettant de déterminer le comportement des contraintes préférentielles, mais cela reste une hypothèse plausible.

Ainsi, la notion de contrainte préférentielle n'est pas identique à celle de Keller (2000) et Sorace & Keller (2005) : elle n'est pas définie en fonction des jugements d'acceptabilité qu'elles induisent et peut varier de langue à langue.

4. « *we predict that there are no constraints that are soft in one language and hard in another* » (Sorace & Keller, 2005, p. 1513).

5. Nous étayerons cette idée dans la section 1.1.2.2 de ce chapitre.

1.1.2. Arguments en faveur de l'intégration des contraintes préférentielles dans le domaine de la syntaxe

Les contraintes préférentielles agissent sur l'acceptabilité des phrases. Or, traditionnellement, la syntaxe ne s'intéresse qu'aux contraintes affectant la grammaticalité, et de ce fait considérées comme les seules à intervenir dans la compétence des locuteurs. L'acceptabilité, quant à elle, concerne des problèmes de performance. Ainsi, la préférence en syntaxe n'est pas l'affaire des grammairiens mais celle des spécialistes de domaines liés à la performance (psycholinguistes, neurolinguistes, sociolinguistes…). Cependant, déterminer quelles sont les contraintes relevant de la compétence et celles relevant de la performance n'est pas une question triviale. Comme le suggère Bresnan (2007a), un détour par les travaux des typologues s'avère très instructif. L'idée centrale est que, dans une langue A, un facteur peut déterminer la grammaticalité d'une construction, alors que, dans une langue B, ce même facteur n'intervient que de façon préférentielle. Le facteur en question doit-il alors être considéré comme une contrainte de compétence dans la langue A et comme une contrainte de performance dans la langue B ?

Nous illustrons ce point en nous appuyant sur deux phénomènes : d'une part, l'interaction de la personne grammaticale et du passif en anglais, en lummi et en picuris (Bresnan *et al.*, 2001) ; d'autre part, l'influence du caractère animé dans la syntaxe des verbes ditransitifs en arménien oriental (Polinsky, 1996), en sesotho (Morolong & Hyman, 1977) et en anglais (Bresnan *et al.*, 2007).

1.1.2.1. Interaction entre personne et voix

Dans certaines langues du monde, l'utilisation de la voix passive et de la voix active est contrainte. Elle est restreinte par la personne grammaticale de l'agent et du patient relatifs au verbe. Pour décrire ce phénomène dans deux langues, nous utiliserons la terminologie suivante : arguments *locaux* pour 1ère et 2ème personne, arguments non-locaux pour 3ème personne, agent et patient du prédicat pour les deux arguments sémantiques d'un verbe transitif.

Zaharlick (1982) observe que, en picuris (langue tanoan, Nouveau Mexique), « *des conditions spécifiques dictent l'utilisation de phrases passives ou actives. À la différence de l'anglais, cette utilisation n'est pas déterminée stylistiquement* »[6]. Elle constate que le passif n'est pas possible quand l'agent et le patient sont tous les deux locaux (4). Si l'agent est local et le patient non-local, le passif est également interdit (6). Enfin, quand l'agent est non-local et que le patient est local, le passif est obligatoire (5).

(4) Agent = 1ère ou 2ème personne / Patient = 1ère ou 2ème personne

a. **ta-mo n-mia-'a n 'e -pa* / **a-mo n-mia-'a n na -pa*
SUJ.1SG-voir-PSF-PSÉ 2SG-par 2SG-voir-PSF-PSÉ 1SG-par

6. « *…specific conditions dictate the use of active or passive sentences. Unlike English, this use is not stylistically determined* » (Zaharlick, 1982, p. 34).

'J'ai été vu par toi' / 'Tu as été vu par moi'

b. *('e) may-mo n-'a n / (na)*
2SG SUJ.2SG+OBJ.1SG-voir-PSÉ 1SG
'a -mo n-'a n
SUJ.1/3SG+OBJ.2SG-voir-PSÉ
'Tu m'as vu' / 'Je t'ai vu'

(5) Agent = 3ème personne / Patient = 2ème personne

a. *'a-mo n-mia-'a n sənene-pa*
SUJ.2SG-voir-PSF-PSÉ homme-par
'J'ai été vu par l'homme' (littéralement)

(6) Agent = 2ème personne / Patient = 3ème personne

a. **sənene mo n-mia-'a n 'e -pa*
homme voir-PSF-PSÉ 2SG-par
'L'homme est vu par toi'

b. *sənene 'a-mo n-'a n*
homme SUJ.2SG-voir-PSÉ
'Tu as vu l'homme'

L'emploi du passif impose donc la présence d'un argument sémantique de 3ème personne et, lorsqu'il y a un argument local et un argument non-local, c'est toujours l'argument local qui doit avoir la fonction sujet par rapport au prédicat.

Le même type de phénomène s'observe en lummi (langue salish, Colombie Britannique) (Jelinek & Demers, 1983, 1994). Dans les propositions principales, si l'agent est local et le patient non-local, le prédicat sera obligatoirement à l'actif (7) ; et inversement, si l'agent est non-local et le patient local, le prédicat sera obligatoirement au passif (8) (exemples tirés de Jelinek & Demers, 1983, p. 168).

(7) Agent = 1ère personne / Patient = 3ème personne

a. *ẋčí-t-sən cə swəy qə*
connaître-TR-1SG le homme
'Je connais l'homme'

b. *____ (*n'existe pas*)
'L'homme est connu par moi'

(8) Agent = 3ème personne / Patient = 1ère personne

a. *____ (*n'existe pas*)
'L'homme me connaît'

b. *ẋčí-t-ŋ-sən ə cə swəy qə*
connaître-TR-PSF-1SG par le homme
'Je suis connu par l'homme'

Ainsi, en lummi comme en picaris, le choix du passif ou de l'actif est contraint par une 'hiérarchie de personne' (*person hierarchy*, Bresnan *et al.*, 2001) selon laquelle les arguments locaux doivent être alignés avec la fonction sujet si l'autre argument est

non-local. Cela signifie que la personne grammaticale est une contrainte catégorique affectant la grammaticalité dans ces deux langues.

En ce qui concerne l'anglais, Bresnan *et al.* (2001) ont observé l'interaction entre la personne grammaticale et la voix passive sur la partie *parsée* du corpus d'anglais américain conversationnel, *Switchboard* (Godfrey *et al.*, 1992). Ils ont extrait les verbes transitifs, relevant pour chacun d'eux la personne grammaticale (locale ou non-locale) de l'agent et du patient, ainsi que la voix (active ou passive) à laquelle est employé le verbe. Sur un total de 10 060 verbes, les auteurs n'observent aucune occurrence de passif, lorsque l'agent est local et quelle que soit la personne du patient. Lorsque l'agent est non-local, Bresnan *et al.* relèvent 2,9% de passifs pour les patients locaux et 1,2% pour les patients non-locaux. Ces observations sont résumées dans la table 1.1.

Agent	Patient	
local	local	0%
local	non-local	0%
non-local	local	2,9%
non-local	non-local	1,2%

TABLE 1.1.: Pourcentages de passifs dans le corpus *Switchboard* selon la personne grammaticale des arguments (Bresnan *et al.*, 2001).

Dans ces données, on ne rencontre la voix passive qu'avec un agent non-local et la proportion de passif est plus importante lorsque le patient est local. Les auteurs en concluent que la contrainte de personne se manifeste par une préférence statistique en anglais.

La comparaison de l'instanciation de la contrainte de personne grammaticale en picuris, en lummi et en anglais montre que le même type de contrainte peut s'exprimer de façon catégorique dans une langue et de façon préférentielle dans une autre. Bresnan *et al.* (2001) estiment que c'est un argument motivant le développement d'un modèle grammatical qui rend compte des *soft constraints*, que nous traduisons par *contraintes préférentielles*.

1.1.2.2. Interaction entre caractère animé des référents et syntaxe des verbes ditransitifs

Dans leur article, Morolong & Hyman (1977) démontrent qu'en sesotho (langue bantoue, Lesotho et Afrique du Sud), le caractère animé peut déterminer la grammaticalité de constructions bénéfactives (ou applicatives). Dans cette langue, il existe un procédé morphologique productif qui permet de passer d'un verbe transitif à un verbe ditransitif dont l'argument supplémentaire est bénéfactif. Le verbe ditransitif est marqué morphologiquement par une 'extension applicative' (*applicative extension*, Morolong & Hyman, 1977). Les arguments thème et bénéfactif apparaissent après le complexe verbal et s'ordonnent selon le caractère humain de leur référent. Si thème

et bénéfactif ont tous les deux un référent humain ou tous les deux un référent non-humain, ils s'ordonnent librement (9) et (10). En revanche, si les deux arguments ont un type de référent différent, c'est l'argument humain qui doit apparaître adjacent au complexe verbal : lorsque le thème (resp. bénéfactif) a un référent humain tandis que le bénéfactif (resp. thème) a un référent non-humain, le premier doit apparaître directement après le complexe verbal (11) et (12) (exemples tirés de Morolong & Hyman, 1977, p. 202-203).

(9) Thème = non-humain / Bénéfactif = non-humain

a. *ke-phehétsé* *mokété lijó*
1SG-cuisiner+APP+PSÉ festin nourriture
'Je cuisine de la nourriture pour le festin'

b. *ke-phehétsé* *lijó* *mokété*
1SG-cuisiner+APP+PSÉ nourriture festin

(10) Thème = humain / Bénéfactif = humain

a. *ke-bítselítsé* *morena baná*
1SG-appeler+APP+PSÉ chef enfant+PL
'J'appelle les enfants pour le chef' ou *'J'appelle le chef pour les enfants'* (ambigu)

b. *ke-bítselítsé* *baná* *morena*
1SG-appeler+APP+PSÉ enfant+PL chef

(11) Thème = non-humain / Bénéfactif = humain

a. *ke-phehétsé* *ngoaná lijó*
1SG-cuisiner+APP+PSÉ enfant nourriture
'Je cuisine de la nourriture pour l'enfant'

b. **ke-phehétsé* *lijó* *ngoaná*
1SG-cuisiner+APP+PSÉ nourriture enfant

(12) Thème = humain / Bénéfactif = non-humain

a. *ke-bítselítsé* *baná* *mokéte*
1SG-appeler+APP+PSÉ enfant+PL festin
'J'appelle les enfants pour le festin

b. **ke-bítselítsé* *mokéte baná*
1SG-appeler+APP+PSÉ festin enfant+PL

En sesotho, le caractère animé, et plus précisément l'opposition humain/non-humain constitue une contrainte catégorique dans l'ordonnancement des compléments de verbes ditransitifs.

La grammaire de l'arménien oriental oral est également sensible au caractère animé des référents : d'après Polinsky (1996), « *le caractère animé est probablement le plus*

important trait catégoriel des noms arméniens »[7]. Dans cette langue, le marquage accusatif est différent selon que le nom est animé ou inanimé. Par exemple, pour l'une des 8 déclinaisons, les noms animés sont marqués d'un suffixe *-i* à l'accusatif, alors que les noms inanimés ne présentent pas de marquage accusatif (Polinsky, 1996, p. 309). Le caractère animé est donc un trait pertinent au niveau morphologique. Ce trait constitue également une contrainte au niveau syntaxique. En effet, il intervient dans l'ordonnancement des compléments des verbes ditransitifs. Polinsky (1996) pose la hiérarchie présentée en (13) pour rendre compte de l'effet du caractère animé en arménien oriental oral.

(13) pronom > nom propre > humain > animal > inanimé[8]
(Polinsky, 1996, p. 329)

Lorsque l'objet direct (OD) et l'objet indirect (OI) d'un verbe ditransitif n'appartiennent pas à la même classe sémantique dans la hiérarchie (13), l'ordre des éléments de la phrase est contraint en fonction de cette hiérarchie. Quand l'OI appartient à une classe supérieure à celle de l'OD, l'OI précède l'OD et le verbe a une position finale (14). Lorsque l'OD appartient à une classe supérieure à celle de l'OI, c'est l'OD qui précède l'OI et le verbe a une position médiane : il apparaît entre l'OD et l'OI (15) (exemples tirés de Polinsky, 1996, p. 329-330).

(14) OD = inanimé / OI = pronom et animal

a. *jes nran hav-ə t'veci*
1SG 3SG.ACC/DAT poulet-ART donner.PSÉ
'Je lui ai donné le poulet'

b. *bžišk'-ə k'at'v-i-n deγ nšanak'ec*
docteur-ART chat-DAT-ART médicament assigner.PSÉ
'Le docteur a prescrit des médicaments pour le chat'

(15) OD = pronom et humain / OI = animal

a. *jes nran t'veci hav-i-n*
1SG 3SG.ACC/DAT donner.PSÉ poulet-DAT-ART
'Je l'ai donné au poulet' (OD = humain)

b. *t'er-ə c'a*Ra-i-n *uγark'ec dzij-u-n*
maître-ART serviteur-ACC/DAT-ART envoyer.PSÉ cheval-DAT-ART
'Le maître a envoyé le serviteur au cheval'

Notons que la contrainte relative à la hiérarchie en (13) neutralise la contrainte de définitude qui intervient lorsque l'OD et l'OI appartiennent à la même classe sémantique. En arménien oriental oral, le caractère animé est donc une catégorie pertinente en morphologie et constitue une contrainte régissant l'ordre des mots dans

7. « *Animacy is probably the most important categorial feature of Armenian nominals* » (Polinsky, 1996, p. 309).
8. A > B signifie que A domine B.

le cas des verbes ditransitifs [9].

Le caractère animé des référents intervient également dans la syntaxe des verbes ditransitifs de l'anglais [10]. Plus précisément, ce facteur intervient dans l'alternance dative. Nous utiliserons la terminologie suivante : le constituant apparaissant comme objet dans la construction à SP datif sera désigné par thème, et nous appellerons destinataire le constituant introduit par la préposition *to* dans la construction à SP datif, comme cela est indiqué dans l'exemple (16).

(16) a. construction à SP datif : *He gives [the picture]$_{theme}$ [to Mary]$_{dest}$*
b. construction à double objet : *He gives [Mary]$_{dest}$ [the picture]$_{theme}$*

L'idée générale est que si le thème, ou le destinataire, est animé, il a tendance à apparaître directement après le verbe. Thompson (1990) a étudié l'influence de plusieurs facteurs (caractère animé, pronominalité, spécificité, définitude, accessibilité…) sur l'alternance dative, à travers une étude sur un corpus de 196 phrases tirées de trois récits écrits. Son corpus révèle notamment que 97% des destinataires sont animés, alors que seuls 5% des thèmes le sont. De plus, parmi les destinataires animés, près de 70% apparaissent dans la construction à double objet, et parmi les six occurrences de destinataires non-animés, cinq apparaissent dans la construction à SP datif. Ces données semblent indiquer qu'un destinataire animé est préférentiellement réalisé comme un double objet adjacent au verbe, tandis qu'un destinataire non-animé a tendance à être réalisé sous forme d'un SP. Plus récemment, d'autres travaux ont confirmé que le caractère animé est une contrainte préférentielle intervenant dans l'alternance dative. En s'appuyant sur l'étude statistique d'un corpus de 2360 observations extraites du corpus *Switchboard* (Godfrey *et al.*, 1992), l'article de Bresnan *et al.* (2007) démontre notamment que la contrainte préférentielle que constitue le caractère animé n'est pas réductible à d'autres facteurs tels que la pronominalité, la complexité du constituant ou l'accessibilité du référent dans le discours. Ces travaux prouvent que le caractère animé est une contrainte qui influence le choix entre les constructions à double objet et à SP datif en anglais [11].

À l'instar de ce qui a été vu pour la contrainte de personne grammaticale, la mise en parallèle de trois langues permet d'observer que le caractère animé constitue une contrainte catégorique en sesotho et en arménien oriental, tandis que son influence sur l'alternance dative en anglais n'est que préférentielle.

Revenons à présent à la différenciation entre contrainte de compétence et contrainte de performance. Rappelons qu'une contrainte catégorique touche à la grammaticalité d'une phrase et fait partie de la compétence des locuteurs. À l'inverse, une contrainte préférentielle en affecte l'acceptabilité et est, à ce titre, considérée comme

9. Certains locuteurs de l'arménien oriental considèrent que cette contrainte ne définit pas l'ordre des compléments du verbe de façon catégorique. Selon eux, cette généralisation ne renverrait qu'à une tendance.

10. L'influence du caractère animé dans les phénomènes touchant à l'ordre des mots sera développée dans la section 5.5.1 du chapitre 5.

11. Nous reviendrons plus en détail sur le travail de Bresnan *et al.* (2007) dans la section 1.3.

une contrainte de performance. Les exemples présentés ci-dessus mènent au constat que seule la façon dont s'exprime la contrainte est différente selon les langues : dans un cas, elle agit de façon catégorique, tandis que dans l'autre, elle ne s'exprime qu'à travers une tendance. Dans la mesure où les contraintes ne relèvent pas de manière uniforme des notions de compétence ou de performance dans toutes les langues, nous formulons l'hypothèse selon laquelle il est difficile de distinguer d'une manière absolue la nature (compétence ou performance) de ces contraintes. Nous supposons qu'il n'existe qu'un seul type de contrainte, qui peut se réaliser de façon catégorique dans une langue et de façon préférentielle dans une autre. Selon cette hypothèse, les contraintes préférentielles font partie de la connaissance qu'ont les locuteurs de leur langue.

Au-delà de l'argument typologique que nous venons de développer, Bresnan & Hay (2008) et Bresnan & Ford (2010) ont montré que l'effet de certaines contraintes préférentielles peut fluctuer entre deux variétés d'une même langue [12].

1.1.2.3. Les contraintes préférentielles dans les différentes variétés de l'anglais

Bresnan & Hay (2008) ont comparé le phénomène de l'alternance dative pour le verbe *give* en anglais américain et en anglais néo-zélandais. Leur travail s'appuie sur une étude de corpus pour ces deux variétés d'anglais. Les résultats montrent que la contrainte préférentielle relative au caractère animé du référent du destinataire a un effet significativement plus important en anglais néo-zélandais qu'en anglais américain. Autrement dit, l'effet de la contrainte caractère animé diffère selon la variété d'anglais considérée. D'après les corpus, un destinataire non-animé a plus de chance d'apparaître dans une construction à double objet en anglais néo-zélandais qu'en anglais américain parlé. Les auteurs concluent que la variabilité des contraintes préférentielles est un argument en faveur de leur intégration dans ce qui constitue la compétence linguistique des locuteurs. En effet, si l'on estime que ces contraintes relèvent simplement de la performance, il faudrait admettre que les processus cognitifs influençant la production sont différents pour les Néo-zélandais et pour les Américains, proposition pour laquelle il n'existe aucune preuve.

Dans le même ordre d'idées, Bresnan & Ford (2010) ont étudié le phénomène de l'alternance dative en anglais américain et en anglais australien, à travers une étude sur corpus et trois études corrélationnelles [13]. Leur article montre notamment que les locuteurs australiens sont plus sensibles à la contrainte de longueur relative du thème et du destinataire que les locuteurs américains. Quand le destinataire est plus long que le thème, la tendance générale est de placer le thème avant le destinataire, c'est-à-dire à produire une construction à SP datif (V SN_{theme} SP_{dest}). Cette préférence est plus marquée chez les locuteurs australiens. Cela s'observe sur corpus (production) et est confirmé par les études corrélationnelles (compréhension et production).

La longueur relative de deux éléments est considérée comme une contrainte de performance puisqu'elle met en jeu des difficultés de traitement et d'analyse chez

12. Nous reviendrons plus en détail sur ces deux travaux dans la partie 1.3.4.
13. L'une de ces trois études corrélationnelles a été présentée dans la section 1.1.1.

les locuteurs (voir par exemple Hawkins, 1994). Comme l'impact de la longueur est plus fort chez les Australiens que chez les Américains, on peut supposer que la connaissance des locuteurs de chacune des deux variétés est subtilement différente à l'égard de cette contrainte préférentielle. Cela constitue un argument de plus en faveur de l'idée que les contraintes préférentielles font partie de la connaissance langagière des locuteurs.

1.1.3. Pourquoi étudier les contraintes préférentielles ?

Approfondir l'étude et la connaissance du langage

En nous appuyant sur les exemples développés dans la partie précédente, nous avons émis l'hypothèse selon laquelle les contraintes préférentielles sont de même nature que les contraintes catégoriques et qu'elles s'expriment de façon diverse dans des langues différentes ainsi qu'à travers les variétés d'une même langue. L'ensemble de ces contraintes, ainsi que leur mode d'expression et leur degré d'importance font partie de la connaissance qu'ont les locuteurs de leur langue. Nous considérons que, tout comme les contraintes catégoriques, les contraintes préférentielles constituent un ensemble de règles à prendre en compte pour expliquer le système de la langue. La grammaire générative et ses méthodes ont permis la description et la formalisation de nombreux phénomènes linguistiques, en s'appuyant principalement sur la mise à jour des contraintes catégoriques. L'approche en termes de contraintes préférentielles permet de compléter la description et la connaissance de phénomènes syntaxiques, là où la grammaire générative ne dit rien.

L'un des intérêts majeurs de l'étude des contraintes préférentielles est l'approfondissement de la description de phénomènes linguistiques afin d'obtenir une meilleure connaissance de la langue. Manning (2003, p. 297) exprime cela de la façon suivante : « *Les théories linguistiques catégoriques expliquent trop peu. Elles ne disent rien du tout à propos des contraintes souples [préférentielles] qui expliquent comment les gens choisissent de dire les choses (ou comment ils choisissent de les comprendre)* » [14]. Si l'on n'étudie pas les contraintes préférentielles, on perd une part importante des généralisations possibles sur des phénomènes linguistiques. Wasow (2002) prend l'exemple de l'ordre des constituants post-verbaux en anglais et du *Principle of End Weight* : « *Le 'Principle of End Weight' qui dépend de mesures relatives de longueur et de complexité, se manifeste paradigmatiquement dans des phénomènes grammaticaux : l'ordre canonique des constituants en anglais (et dans d'autres langues), et le caractère obligatoire d'une forme dans certaines alternances syntaxiques quand l'objet direct est un pronom. Si l'ordre canonique et les cas obligatoires font partie de la grammaire de compétence tandis que les préférences quantitatives sont traitées comme de la performance, alors une généralisation plus large est perdue* » [15]. Le principal intérêt de l'étude des contraintes préférentielles est identique

14. « *Categorical linguistic theories explain too little. They say nothing at all about the soft constraints which explain how people choose to say things (or how they choose to understand them).* »

15. « *the Principle of End Weight, which depends on relative measures of length and complexity,*

à celui de la syntaxe en général : décrire et comprendre le système de la langue. La recherche autour de la formalisation des contraintes préférentielles et de leur interaction constitue également un moyen d'améliorer la connaissance de l'organisation des propriétés de la langue.

Intérêt typologique

L'étude des contraintes préférentielles présente un deuxième intérêt, à savoir la comparaison des langues. Comme nous l'avons vu dans les sections 1.1.2.1 et 1.1.2.2, les contraintes linguistiques s'expriment différemment selon les langues : une contrainte catégorique dans une langue A peut être préférentielle dans une langue B. La comparaison des contraintes à travers les langues passe par l'étude des contraintes préférentielles dans chaque langue, prise individuellement.

Applications

Nous voyons un troisième aspect qui révèle l'intérêt de ce type d'étude : les applications telles que la génération automatique, l'analyse automatique de textes ou l'acquisition d'une langue seconde. En effet, la mise à jour des contraintes préférentielles intervenant dans une langue, ainsi que la formalisation de leur intervention dans un phénomène peuvent permettre, à terme, d'améliorer ces applications. En ce qui concerne la génération, l'utilisation de contraintes préférentielles pourrait guider le choix d'une structure ou d'un ordre, rendant le texte plus naturel. Les analyseurs syntaxiques statistiques contemporains captent probablement des préférences du type de celles qui nous intéressent, mais de façon non-explicite. L'étude des contraintes préférentielles dans une perspective linguistique pourrait donc représenter, à terme, un moyen de mieux comprendre la "boîte noire" que constitue un analyseur statistique. Enfin, dans un tout autre domaine, l'apprentissage de langues étrangères, les travaux sur les contraintes préférentielles et leur interaction pourraient aider les professeurs à expliquer des phénomènes de préférences qui sont souvent difficiles à enseigner, par exemple le choix de la construction à double objet ou à SP datif en anglais, ou bien la position de l'adjectif épithète par rapport au nom en français.

L'étude des contraintes préférentielles impose une réflexion méthodologique, dans la mesure où la méthode traditionnelle, qui vise à classer les séquences selon leur grammaticalité, n'est pas envisageable. Les outils utilisables, en particulier les corpus annotés et les expériences psycholinguistiques, posent des problèmes de généralisation et de formalisation des aspects non-catégoriques du système de la langue, que nous abordons dans la section suivante.

---

manifests itself in paradigmatically grammatical phenomena : the canonical constituent order of English (and other languages), and the obligatoriness of one form in certain syntactic alternations when the direct objet is a pronoun. If the canonical ordering and the obligatory cases are part of the competence grammar, but the quantitative preferences are treated as performance, then a larger generalization is lost » (Wasow, 2002, p. 139).

1.2. Méthodes et généralisation

Nous envisageons trois outils méthodologiques pour l'étude des contraintes préférentielles : l'introspection et les jugements de grammaticalité, les corpus annotés ainsi que les expériences psycholinguistiques. Ces outils sont complémentaires, dans la mesure où ils permettent d'étudier différentes facettes du système de la langue et que chacun d'entre eux permet de dépasser les limites des autres outils.

1.2.1. Introspection et jugement de grammaticalité

L'utilisation du jugement de grammaticalité, basé sur l'introspection, repose sur une distinction nette entre compétence et performance. Cette méthode constitue la source essentielle de données dans la très grande majorité des travaux syntaxiques. Elle fait appel à l'intuition de locuteurs natifs à propos de l'appartenance de phrases au système de la langue. Il est important de préciser le statut du jugement de grammaticalité : les locuteurs natifs ne fournissent pas des jugements de grammaticalité, dans la mesure où la notion de grammaticalité est une construction théorique. En effet, les intuitions des locuteurs ne sont que des jugements d'acceptabilité et c'est au linguiste de décider si ces jugements sont le reflet d'une propriété de la compétence ou de phénomènes liés à la performance. Par exemple, Newmeyer (1983) écrit : « *puisque la grammaticalité est une construction théorique, elle n'est pas directement accessible aux intuitions du locuteur de la langue. Un locuteur n'a pas plus d'intuition sur la grammaticalité d'une phrase que sur le fait que la phrase "John gave Sue the book" est formée grâce à une règle lexicale ou une règle transformationnelle, ou sur le fait que le niveau approprié pour spécifier les conditions régissant les relations d'anaphore avec l'antécédent est la structure de surface* » [16]. Selon cet auteur, les locuteurs ne peuvent avoir des avis que sur l'acceptabilité des phrases et seuls les linguistes, en tant que professionnels spécialisés, peuvent avoir des intuitions sur la grammaticalité, mais cela reste du domaine des constructions théoriques. Schütze (1996), pour sa part, estime que « *cela n'a aucun sens de parler de jugements de grammaticalité étant donné les définitions de Chomsky, parce que les gens sont incapables de juger la grammaticalité – ce n'est pas accessible à leurs intuitions* » [17]. Il faut établir le statut grammatical d'une phrase à partir de jugements d'acceptabilité. Ainsi, si une phrase est jugée inacceptable, il faut déterminer si cela relève de la performance ou de la compétence. Dans le premier cas, on émet l'hypothèse que la phrase est grammaticale ; dans le deuxième, on conclut que la phrase est agrammaticale. Les

16. « *Since grammaticality [...] is a theoretical construct, it is not directly accessible to the intuitions of the speaker of the language. A speaker no more has intuitions about the grammaticality of a sentence than about whether the sentence "John gave Sue the book" is formed by a lexical rule or a transformational rule, or whether the proper level to state conditions governing antecedent anaphor relations is surface structure.* » (Newmeyer, 1983, p. 50-51).

17. « *It does not make any sense to speak of grammaticality judgements given Chomsky's definitions, because people are incapable of judging grammaticality – it is not accessible to their intuitions.* » (Schütze, 1996, p. 26).

intuitions des locuteurs sont donc des jugements d'acceptabilité et les jugements de grammaticalité correspondent à une décision théorique du grammairien.

La démarche reposant sur les intuitions et les jugements de grammaticalité est hypothético-déductive, dans la mesure où elle consiste à émettre une hypothèse générale permettant d'expliquer les propriétés observées. Ces propriétés sont notamment les contrastes entre phrases jugées acceptables et inacceptables [18], que le chercheur interprète comme des contrastes de grammaticalité. Ce dernier formule alors une (ou plusieurs) hypothèse(s) explicative(s) permettant de saisir ces contrastes. La description du système de la langue repose donc sur la formulation d'hypothèses explicatives rendant compte des jugements de grammaticalité du linguiste.

L'utilisation de l'intuition des locuteurs et des jugements de grammaticalité pose des problèmes méthodologiques. Nous développons plus en détail deux aspects qui ont fait l'objet de critiques : d'une part, le manque de contrôle des conditions dans lesquelles le jugement est produit et l'absence de véritable protocole expérimental ; d'autre part, les données qui sont soumises aux jugements.

Premièrement, le recueil de l'intuition des locuteurs s'apparente à une expérience. Cependant, à la différence d'expériences menées en laboratoire, aucun protocole expérimental n'est mis en place de façon à contrôler les facteurs extra-linguistiques pouvant avoir une influence sur les jugements. Schütze (1996, p. 52-53) identifie trois types de problèmes relatifs à cette méthode : les jugements de grammaticalité ne sont pas systématiquement reportés et la notation n'est pas unifiée ; certains jugements de grammaticalité sont parfois retenus ou écartés selon les auteurs ; certains jugements de grammaticalité demandent aux locuteurs des capacités spécifiques, sans aucune preuve que ces derniers sont vraiment capables de faire les distinctions demandées et sans aucun contrôle du processus d'obtention des jugements. Ces critiques remettent en cause la validité des données issues des méthodes traditionnelles en linguistique, notamment celles touchant à des données controversées.

Deuxièmement, sont soumises au jugement des phrases construites et décontextualisées. Cela peut conduire au rejet de certaines structures qui pourraient pourtant être produites dans des contextes plus élaborés (Bresnan & Nikitina, 2009; Lødrup, 2007; Manning, 2003; Riemer, 2009; Taylor, 1996). À ce propos, plusieurs travaux ont montré comment les jugements de grammaticalité rencontrés dans les travaux de syntaxe « *sous-estiment l'espace de possibilité grammaticale* » [19] selon les mots de Bresnan (2007a) [20]. Par exemple, Manning (2003) observe la notion de cadre de sous-

18. Nous simplifions ici en réduisant les jugements à une catégorisation binaire, alors qu'il peut exister une gradation.

19. « *…underestimate the space of grammatical possibility …* » (Bresnan, 2007a, p. 1)

20. Du point de vue interne à la grammaire générative, le décalage entre les jugements de grammaticalité rencontrés dans les travaux de syntaxe et les constructions effectivement attestées ou jugées acceptables par un ensemble de locuteurs ne remet pas en cause les données utilisées, dans la mesure où, d'un point de vue théorique, grammaticalité et acceptabilité ne sont pas deux notions équivalentes. On s'attend donc à ce que jugements de grammaticalité et jugements d'acceptabilité ne coïncident pas : le fait qu'une phrase soit inacceptable ne prouve pas qu'elle soit agrammaticale et, inversement, le fait qu'une phrase soit acceptable ne prouve pas qu'elle soit grammaticale. Cependant, comme l'explique Riemer (2009), le décalage entre données attestées et jugements de

catégorisation à la lumière de données de corpus. Les cadres de sous-catégorisation que l'on trouve dans les travaux de grammaire générative reposent sur des jugements de grammaticalité ne reflètant pas l'usage. L'auteur cite les jugements de Pollard & Sag (1994) pour les verbes *consider*, *regard*, *turn out* et *end up* et les confronte à des phrases issues du *New-York Times*. Nous reproduisons ici un exemple de phrase jugée agrammaticale suivie d'un contre-exemple du *New-York Times*.

(17) a. **We* ***consider*** *Kim* ***as*** *an acceptable candidate*
b. "*The boys* ***consider*** *her* ***as*** *family and she participates in everything we do*"

(18) a. **We* ***regard*** *Kim* ***to be*** *an acceptable candidate.*
b. "*Conservatives argue that the Bible* ***regards*** *homosexuality* ***to be*** *a sin.*"

(19) a. **Kim* ***turn out doing*** *all the work*
b. "*But it* ***turned out*** *hav****ing*** *a greater impact than any of us dreamed*"

(20) a. **Kim* ***ended up*** *sen****t*** *more and more leaflets.*
b. "*On the big night, Horatio* ***ended up*** *flatten****ed*** *on the ground like a fried egg with the yolk broken*"

Les contre-exemples présentés ici ne sont pas des fautes ayant échappé aux relecteurs du journal. En effet, on trouve dans le journal plusieurs autres contre-exemples de ce type et les locuteurs de l'anglais semblent juger ces phrases acceptables.

Dans le même ordre d'idées, Bresnan & Nikitina (2009) montrent, à propos de l'alternance dative en anglais (cf. exemples (16)), que des constructions jugées agrammaticales peuvent se rencontrer en corpus. Les auteurs confrontent des jugements de grammaticalité rencontrés dans des travaux de linguistique à ce que l'on peut trouver dans les corpus. Beaucoup de travaux (Davidse, 1996; Georgia, 1974; Goldberg, 1995; Gropen *et al.*, 1989; Krifka, 2004; Levin, 1993; Oehrle, 1976; Pinker, 1989) ont étudié les raisons et les possibilités d'alternance. Bresnan & Nikitina (2009) reprennent quatre affirmations issues de ces travaux :

1. les verbes exprimant la transmission continue d'une force apparaissent uniquement dans la construction à SP datif :

   (21) **I lowered John the box/ I lowered the box to John.*

grammaticalité est problématique dans la perspective du pouvoir explicatif de la théorie. En effet, l'un des objectifs d'une théorie scientifique est d'être capable d'expliquer, ou de *prédire*, les faits observés. Or, le seul moyen de vérifier si les prédictions de grammaticalité faites par la théorie sont correctes est d'émettre l'hypothèse selon laquelle l'ensemble des phrases grammaticales et l'ensemble des phrases acceptables coïncident dans la très grande majorité des cas, lorsqu'aucun facteur de performance n'intervient. Plus une théorie est capable de prédire les jugements d'acceptabilité des locuteurs, plus elle a un pouvoir prédictif important et plus elle est satisfaisante. Le grand décalage observé entre les jugements de grammaticalité et les jugements d'acceptabilité remet donc en cause le pouvoir explicatif des théories prédisant ces jugements de grammaticalité. Riemer (2009) affirme que « *If these assignments are as often wrong as the theory's detractors claim, there is a serious empirical deficit in the theory's predictive and explanatory power* » (Riemer, 2009, p. 630).

2. Les verbes exprimant une manière de parler n'apparaissent que dans la construction à SP datif :

   (22) **Susan muttered Rachel the news/ Susan muttered the news to Rachel.*

3. Certains emplois idiomatiques de *give* ne peuvent pas apparaître dans des constructions à SP datif :

   (23) *The noise gave Terry a headache/ *The noise gave a headache to Terry.*

4. Les verbes *cost* et *deny* apparaissent uniquement avec la construction à double objet :

   (24) *The car cost Beth 5000$/ *The car cost 5000$ to Beth*

Pour chacun de ces postulats, les auteurs livrent des contre-exemples trouvés en corpus. Nous reproduisons ici un contre-exemple par catégorie :

1. les verbes exprimant la transmission continue d'une force :

   (25) *Therefore, when he got to purgatory, Buddha* ***lowered him the silver thread of a spider*** *as his last chance for salvation.*

2. Les verbes exprimant une manière de parler :

   (26) *Shooting the Urasian a surprised look, she* ***muttered him a hurry apology*** *as well before skirting down the hall.*

3. Certains emplois idiomatiques de *give* :

   (27) *From the heads, offal and the accumulation of fishy, slimy matter, a stench or smell id diffused over the ship that would give* ***a headache to the most athletic constitution***

4. Les verbes *cost* et *deny* :

   (28) *He did so thinking it would* ***cost nothing to the government***

Ces exemples montrent que les contraintes lexicales établies dans la littérature ne sont pas des contraintes catégoriques. Certains contextes autorisent l'emploi de verbes avec des constructions qui, hors contexte, ne semblent pas possibles. Les exemples de la sous-catégorisation et de l'alternance dative en anglais montrent que l'acceptabilité de certaines constructions n'est pas bien représentée par les jugements de grammaticalité d'exemples construits et décontextualisés.

Dans la suite de cette section, nous montrons que l'étude de corpus annotés, ainsi que l'utilisation d'expériences psycholinguistiques constituent des moyens de dépasser

les deux critiques formulées à l'encontre de la méthodologie employée traditionnellement en syntaxe. Nous explicitons également en quoi ces deux méthodes sont utiles pour l'étude des préférences en syntaxe.

1.2.2. Corpus annotés

L'utilisation de corpus annotés est une méthode complémentaire à l'introspection. Elle permet notamment de surmonter les limitations dues aux phrases construites et décontextualisées, et de prendre en compte l'usage de la langue. En utilisant des données attestées, le chercheur prend en considération des constructions effectivement produites, et ce, dans des conditions non-expérimentales. Il est important de noter que l'annotation en corpus, notamment les parties du discours, les structures et les dépendances syntaxiques, sont elles-mêmes issues de la démarche introspective et des jugements de grammaticalité. L'utilisation de corpus annotés revient à utiliser une première "couche" de connaissances afin de développer l'étude d'autres phénomènes.

Les données de corpus permettent d'observer des tendances que nous interprétons comme des préférences. Nous émettons une hypothèse forte selon laquelle les observations quantitatives sur corpus sont en correspondance avec une forme de savoir langagier. Cette hypothèse n'est plausible que dans le cas de phénomènes stables dans la langue et non en cours d'évolution.

En plus de l'observation des contraintes préférentielles, l'analyse de données de corpus permet d'observer l'interaction de ces contraintes pour un phénomène donné. En effet, les données attestées en contexte mettent en jeu simultanément des contraintes multiples. Une analyse adéquate de ces données permet d'évaluer l'interaction et l'importance relative des contraintes mises en jeu. Par exemple, dans leur travaux sur l'alternance dative en anglais (cf. exemple (16)), Gries (2003b) et Bresnan *et al.* (2007) ont développé une analyse statistique à partir de données de corpus[21]. Ils ont notamment montré que l'alternance dative est un phénomène multi-factoriel guidé par un ensemble de contraintes non-catégoriques. D'après leurs travaux, les contraintes influençant le choix de la construction à double objet ou à SP datif sont :

- la classe sémantique du verbe (Bresnan *et al.*, 2007) / le verbe exprime un transfert ou non (Gries, 2003b)
- l'accessibilité du destinataire et du thème (Bresnan *et al.*, 2007; Gries, 2003b)
- la pronominalité du destinataire et du thème(Bresnan *et al.*, 2007; Gries, 2003b)
- la définitude du destinataire et du thème (Bresnan *et al.*, 2007; Gries, 2003b)
- le caractère animé du destinataire et du thème (Bresnan *et al.*, 2007; Gries, 2003b)
- la longueur du destinataire et du thème (Gries, 2003b) / différence de longueur entre le thème et le destinataire (Bresnan *et al.*, 2007)
- le nombre du destinataire et du thème (Bresnan *et al.*, 2007)
- la personne grammaticale du destinataire (Bresnan *et al.*, 2007)

De façon générale, les éléments accessibles, pronominaux, définis et animés ont

21. Ces deux travaux seront décrits plus en détail dans la partie 1.3.3.

tendance à apparaître adjacents au verbe. Ainsi, si le destinataire a ces qualités, la construction aura tendance à être à double objet (V $\mathbf{SN}_{dest}$ SN), et inversement, si c'est le thème qui présente ces qualités, on aura plutôt la construction à SP datif (V $\mathbf{SN}_{theme}$ SP). Étant donné la variation possible des valeurs de ces variables pour le destinataire et pour le thème, on observe une interaction complexe que Gries (2003b) et Bresnan *et al.* (2007) modélisent à l'aide d'outils statistiques. Seule une approche sur corpus permet d'étudier de telles interactions. En effet, dans le cadre de méthodes se fondant sur les jugements de grammaticalité, l'objectif est de créer des contrastes permettant d'isoler une seule contrainte et de pouvoir ainsi tirer des conclusions sur son effet. L'utilisation de corpus annotés accompagnés de méthodes d'analyse de données adéquates permet d'apporter des observations nouvelles et complémentaires pour décrire et comprendre un ensemble de phénomènes syntaxiques.

L'utilisation des données attestées, ainsi que l'hypothèse de l'existence d'une correspondance entre les tendances observées en corpus et les préférences des locuteurs soulèvent d'importants problèmes méthodologiques : la question de la généralisation, le problème de la prise en compte de données attestées pouvant être jugées comme agrammaticales et les difficultés liées aux corrélations des contraintes étudiées.

La question de la généralisation peut être formulé de la façon suivante : comment passer d'observations dans des corpus à des propriétés générales de la langue ? À la différence des jugements de grammaticalité, les données attestées ne comportent pas d'exemple négatif permettant de dresser des contrastes que l'on tente d'expliquer en formulant des hypothèses sur le système de la langue. La méthode envisagée pour obtenir des propriétés générales sur la langue est l'analyse de données statistique. La démarche sous-jacente à ce type d'analyse est contraire à l'approche hypothético-déductive liée à l'utilisation des jugements de grammaticalité. En effet, il s'agit d'induire les propriétés de la langue à partir d'un ensemble fini d'observations. Nous exposerons en détail les méthodes statistiques ainsi que les hypothèses de travail qu'elles impliquent dans le chapitre 2. Dans la présente section, nous tentons de donner des arguments pour répondre à une question plus théorique : comment passer d'observations effectuées sur une communauté de locuteurs à la connaissance détenue par le locuteur individuel ? Autrement dit, en quoi les fréquences observées en corpus sont en rapport avec la production d'un locuteur ? Pour tenter de répondre à cette question, nous émettons la conjecture selon laquelle l'usage façonne, au moins en partie, la connaissance des locuteurs (Bybee, 2006, 2010; Croft, 2001; Goldberg, 2006) [22]. Dans cette perspective, le corpus est conçu comme un échantillon représentatif de la langue. Cette hypothèse quant à la représentativité du corpus pose un problème non résolu que nous aborderons en détail dans le chapitre 2. Les fréquences des constructions qui y sont observées sont donc représentatives de la fréquence de ces constructions dans la langue. Chaque locuteur est soumis à des fréquences similaires à celles rencontrées dans l'échantillon représentatif que constitue le corpus.

22. Notons que l'usage est considéré comme pertinent dans des domaines linguistiques autres que la syntaxe. Par exemple, le phénomène de la productivité morphologique est partiellement expliqué en termes de fréquence par Baayen (2003).

Si la réprésentation linguistique des locuteurs est effectivement affectée par la fréquence d'occurrences des constructions, alors l'étude des constructions en corpus en fonction de leur fréquence permet d'induire des connaissances sur le savoir langagier des locuteurs. De plus, certains travaux ont montré l'existence d'un lien entre les préférences calculées à partir de données de corpus et les jugements portés par les locuteurs. Par exemple, Bresnan (2007b) a observé, dans le cas de l'alternance dative, que la probabilité d'avoir la construction à double objet (V SN SN), estimée à partir des données de corpus, présentait une corrélation significative avec les préférences exprimées par les locuteurs pour cette même construction. L'auteur a également montré que les jugements des locuteurs pouvaient en grande partie s'expliquer en fonction de contraintes identiques à celles identifiées en corpus. Plus précisément, la plupart des contraintes préférentielles influençant la probabilité d'occurrences de la construction à double objet sont identifiées comme influençant significativement le jugement des locuteurs sur cette construction. Les jugements des locuteurs peuvent alors être analysés comme le reflet de la probabilité d'occurrence de la construction en présence des différentes contraintes préférentielles. Selon les mots de Bresnan, « *les utilisateurs de la langue peuvent faire des prédictions probabilistes précises des choix syntaxiques des autres* » [23]. De la même façon que la grammaire générative a postulé un lien entre jugements de grammaticalité et contraintes syntaxiques catégoriques, ce type de résultat permet de postuler l'existence d'un lien entre contraintes préférentielles et jugements d'acceptabilité des locuteurs. Notons, cependant, que jugements et contraintes n'entretiennent pas le même type de rapport dans ces deux cas : dans le premier cas, les jugements sont des observations et les règles syntaxiques sont des hypothèses formulées pour rendre compte de ces observations ; dans le deuxième cas, les contraintes préférentielles sont issues de l'observation et de l'analyse quantitative de données attestées et les jugements sont d'autres observations obtenues indépendamment. La correspondance entre jugements et contraintes préférentielles est ensuite établie en raison de l'influence des mêmes contraintes en corpus et sur les jugements.

Le deuxième obstacle à l'utilisation de données attestées réside dans le fait que les corpus contiennent parfois des énoncés qui peuvent être jugés agrammaticaux. Se pose alors la question de savoir comment construire une représentation du système de la langue à partir de données contenant des exemples ne répondant pas aux règles de ce système. Premièrement, si la quantité d'énoncés jugés agrammaticaux est marginale, leur influence sur les conclusions d'un travail quantitatif se fondant sur des méthodes d'analyse robustes ne sera pas significative. Les prendre en compte, en respectant les proportions qu'elles représentent, n'apparaît pas réellement problématique dans la perspective quantitative qui nous intéresse.

L'analyse de données attestées comporte un troisième problème : la corrélation des contraintes. Dans la mesure où les données de corpus ne doivent pas être manipulées, les corrélations ne peuvent pas être contrôlées. Par exemple, les constituants longs

23. « *language users can in effect make accurate probabilistic predictions of the syntactic choices of others* » (Bresnan, 2007b, p. 91).

ont tendance à être syntaxiquement complexes, et inversement. Il est donc difficile, à partir de données de corpus, de démêler l'effet de la longueur de celui de la complexité. En revanche, un travail à partir de données construites permet de contrôler ce type de corrélations. Il est, par exemple, possible de construire des énoncés où la longueur d'un constituant est constante mais où sa complexité syntaxique varie. La possibilité de décorréler les contraintes est un des aspects des méthodes expérimentales constituant un complément à l'analyse des données de corpus.

1.2.3. Expériences psycholinguistiques

Parmi les nombreuses possibilités d'expériences psycholinguistiques que l'on peut mener dans le cadre d'un travail de syntaxe, nous nous concentrons sur l'élicitation de jugement d'acceptabilité à l'aide de questionnaires.

Comme nous l'avons mentionné précédemment, le recueil de jugements d'acceptabilité, tel qu'il est pratiqué traditionnellement en syntaxe, est problématique dans la mesure où le protocole mis en place n'est pas assez contrôlé et laisse place à d'importants biais expérimentaux. Schütze (1996) et Cowart (1997) plaident pour une rénovation des méthodes de collecte de données et proposent des protocoles expérimentaux pour les linguistes. Ils suggèrent de caractériser précisément l'acte de jugement et présentent une série de précautions méthodologiques destinées à contrôler les effets extra-linguistiques de la tâche de jugement : par exemple, nombre de participants contrôlé, non connaissance du phénomène étudié par les locuteurs interrogés, randomisation des phrases pour éviter les effets de persistance [24], instructions claires et précises, réflexion sur l'échelle de grammaticalité proposée aux locuteurs…

À partir d'un protocole expérimental respectant les principes méthodologiques de la psycholinguistique, il est possible de tester l'effet d'une contrainte sur les jugements concernant un phénomène spécifique [25]. Les données expérimentales sont construites de façon à contrôler l'ensemble des contraintes, autres que celle étudiée, qui pourrait influencer les préférences des sujets et ainsi permettre d'imputer à la seule contrainte qui varie les différences observées dans les jugements.

Cette méthode constitue un complément à l'utilisation de corpus annotés car elle permet de décorréler les variables. Il est également possible d'utiliser des questionnaires psycholinguistiques dans le but de vérifier la plausibilité des préférences observées sur corpus. C'est ce que fait Bresnan (2007b) dans le travail que nous avons évoqué dans la section précédente.

Les trois méthodes mentionnées sont complémentaires. L'analyse des jugements des locuteurs en termes de grammaticalité permet de tracer l'architecture d'une langue en définissant les contraintes catégoriques qui structurent son système. Les deux autres méthodes permettent l'étude et la mise à jour de contraintes préférentielles qui

24. Traduction du terme anglais *priming*. Pour plus de détails sur la notion d'effet de persistance, voir section (25), chapitre 6.

25. Pour un exemple de mise en pratique de ce type de méthodes en syntaxe, voir Keller (2000).

agissent dans les zones de liberté offertes par la structure syntaxique fondamentale de la langue. L'étude de phénomènes préférentiels prend donc appui sur le savoir défini en termes catégoriques, afin d'approfondir l'étude du langage. Nous nous intéressons plus particulièrement aux deux dernières méthodes, car ce sont elles qui permettent d'aborder notre objet d'étude. Elles seront traitées dans le chapitre 2.

1.3. L'exemple de l'alternance dative

L'alternance dative en anglais a fait l'objet de nombreux travaux, dont certains proposent une analyse compatible avec les contraintes préférentielles et reposent sur les méthodes qui nous intéressent. Nous présentons en détail les travaux concernant ce sujet, dans le but de montrer ce que l'étude des contraintes préférentielles peut apporter à la description et à la compréhension de l'alternance dative et comment les appréhender de façon satisfaisante.

Rappelons que la construction dative en anglais met en jeu un thème et un destinataire, pouvant être réalisés sous la forme d'une construction à SP datif (V SN_{theme} SP_{dest}) ou bien d'une construction à double objet (V SN_{dest} SN_{theme}), comme cela est montré dans l'exemple (29).

(29) a. construction à SP datif : *Kim handed a toy to the baby*
b. construction à double objet : *Kim handed the baby a toy*

1.3.1. Contre une explication purement sémantique

La littérature sur le problème de l'alternance dative en anglais est très étendue. Une part importante de ces travaux fait reposer l'alternance sur une différence de sens. Plus précisément, la construction à double objet est associée à un sens de possession et exprime un évènement causant un changement d'état (30-a), alors que la construction à SP datif est liée à un sens de mouvement qui reflète le changement de place, le mouvement vers un but (30-b) (Bresnan *et al.*, 2007).

(30) a. Possession : '**x** entraîne que **y** possède **z**' ⇒ V SN SN
b. Mouvement : '**x** entraîne que **z** va vers/est à **y**' ⇒ V SN SP
Tiré de Levin & Rappaport Hovav (2002, p 1).

Les arguments en faveur de cette explication sémantique sont notamment l'impossibilité d'avoir la construction à double objet avec un verbe exprimant la transmission continue d'une force (cf. (21)) ; ou bien avec un verbe exprimant une manière de parler (cf. (22)) ; ou encore avec les emplois idiomatiques de *give* (cf. (23)). Comme nous l'avons vu dans la section 1.2.1, ces arguments présentent des contre-exemples, ce qui remet en cause l'explication purement sémantique de l'alternance. Cependant, il est vrai qu'il existe une tendance à exprimer le transfert de possession avec la construction à double objet. Bresnan & Nikitina (2009) proposent une explication à cette tendance. Selon ces auteurs, le transfert de possession peut être

exprimé par les deux constructions. Or, le verbe *give*, qui est le verbe prototypique du transfert de possession, apparaît dans la très large majorité des cas avec la constructions à double objet (84.6% dans le corpus *Switchboard*). Par conséquent, l'expression du transfert de possession est généralement associée à la construction à double objet. Par ailleurs, certains verbes (transmission continue d'une force, manière de parler) décrivent rarement des situations où il y a transfert de possession. Donc, ces deux classes de verbes apparaissent très fréquemment avec la construction à SP datif et très rarement avec le double objet. Les exemples avec le double objet sont alors jugés agrammaticaux hors contexte car ils ne renvoient pas à une situation souvent décrite.

Étant donné que les deux constructions peuvent exprimer les sens de possession et de mouvement, la sémantique des verbes n'est pas une explication suffisante ; elle ne représente qu'une contrainte préférentielle en interaction avec d'autres.

1.3.2. Études sur corpus

Le corpus constitue un accès privilégié à l'étude de l'altenance dative, dans la mesure où un nombre important de contraintes préférentielles intervient dans ce phénomène. Dans cette section, nous évoquerons trois études qui traitent de ce sujet à partir de données de corpus : Thompson (1990), Collins (1995) et Snyder (2003).

L'article de Thompson (1990) s'intéresse à la caractérisation des destinataires, du point de vue de la notion de *topic*. Le corpus utilisé est composé de trois récits écrits [26]. Cent quatre-vingt seize phrases en ont été extraites, et 71% d'entre elles présentent des constructions à double objet. Ces phrases comportent toutes un verbe ditransitif accompagné d'un thème et d'un destinataire postverbaux et, pour chacune d'elles, l'ordre alternatif est considéré sémantiquement acceptable par l'auteur. Cette dernière utilise le concept de *topicworthiness*, qu'elle définit comme un ensemble de propriétés influençant la probabilité qu'un SN soit le *topic* d'un discours. Les propriétés étudiées sont au nombre de sept :

1. caractère animé,
2. pronominalité,
3. spécificité du référent,
4. identifiabilité du référent,
5. nom propre,
6. état d'activation du référent dans la conscience du destinataire,
7. longueur en nombre de mots.

L'auteur montre que les destinataires présentent significativement plus de propriétés renvoyant au *topic* que les thèmes. Elle observe également que les destinataires apparaissant dans la construction V $\mathbf{SN}_{dest}$ SN ont plus de caractéristiques renvoyant au *topic* que ceux qui apparaissent dans la construction V SN $\mathbf{SP}_{dest}$. Ces résultats s'appuient sur des observations de fréquence et des tests de χ^2. [27]

26. *Murder at the Vicarage* de Agatha Christie ; *Have His Carcase* de Dorothy Sayres et *Nim* de Herbert Terrace.

27. Pour avoir des précisions sur le test de χ^2, se reporter au chapitre 6 de Howell (1998).

Le travail de Collins (1995) s'inspire de celui de Thompson (1990). Il cherche également à caractériser l'influence des aspects informationnels dans l'alternance dative, à partir d'une étude sur corpus. Cependant, à la différence de l'étude précédemment décrite, Collins étudie la variété de l'anglais australien, en utilisant un corpus mieux échantillonné que celui de Thompson. Ce corpus comporte environ 200 000 mots dont 100 000 mots renvoient à des données de l'oral (conversations pendant des repas, discours parlementaires) et les 100 000 autres correspondent à des textes écrits (biographies, essais). Cent soixante-cinq phrases comportant l'ordre V SN SN ou V SN SP en ont été extraites et 65% d'entre elles présentent la construction à double objet. Quatre variables ont été étudiées, à savoir l'accessibilité, la pronominalité, le caractère défini et la longueur en nombre de mots. L'accessibilité est définie comme une variable servant à capter dans quelle mesure un référent est récupérable à partir du co-texte ou de la situation de communication. Elle se décline selon trois valeurs : 'donné', 'accessible' et 'nouveau' [28]. À partir de relevés de fréquence, l'auteur montre que ces quatre variables se répartissent de façon différente selon que le référent est le destinataire ou le thème et selon que la construction observée est à double objet ou à SP datif. Il observe d'abord que le destinataire présente, plus souvent que le thème, les quatre propriétés décrites, ce qui confirme les observations de Thompson (1990). L'auteur remarque également que, dans la construction à double objet, le destinataire, réalisé comme un SN directement après le verbe, a tendance à regrouper les quatre caractéristiques étudiées, tandis que dans la construction à SP datif il présente moins fréquemment les propriétés étudiées. Il en conclut que la différence de statut informationnelle entre le destinataire et le thème est plus marquée dans la construction à double objet que dans la construction à SP datif. Il considère donc que cette dernière est plus neutre quant au statut informationnel du destinataire.

Le troisième travail sur corpus est celui de Snyder (2003). Cette étude se différencie des deux précédentes par la taille de l'échantillon analysé et par le soin pris à considérer l'item verbal comme une variable pertinente. À partir de corpus d'oral (*Switchboard*, Godfrey *et al.*, 1992) et de corpus écrits (2 romans), Snyder a relevé 1666 phrases présentant les verbes *give*, *sell*, *send*, *bring* et *take* dans une construction à double objet ou à SP datif. À l'instar de Thompson (1990) et de Collins (1995), elle observe le rôle de la longueur et du statut informationnel du destinataire et du thème, en fondant son analyse sur une étude des fréquences.

Les observations effectuées dans les trois travaux décrits précédemment sont des tendances sur corpus. Elles semblent difficiles à considérer comme représentatives de la langue anglaise, ou même d'une de ses variétés, et ce, pour trois raisons principalement. Premièrement, les observations s'appuient sur l'étude d'un certain nombre de variables sélectionnées. Cela implique que l'on néglige l'effet d'autres variables qui ont une influence. Les tendances constatées offrent donc un biais non contrôlé. Deuxièmement, les variables intervenant dans l'alternance dative sont nombreuses et souvent corrélées. Étant donné qu'aucune précaution méthodologique n'est prise pour contrôler les corrélations et pour observer l'effet d'une variable en fonction des

28. Traduits de l'anglais *given*, *accessible* et *new*.

autres, on ne peut pas déterminer si le comportement d'une variable n'est pas réductible à une ou plusieurs autres variables. Troisièmement, les tendances ne sont que des observations de fréquence sur des corpus spécifiques, plus ou moins représentatifs, contenant un nombre réduit d'occurrences (sauf pour le travail de Snyder, 2003). Ces tendances ne peuvent donc pas être généralisées au-delà des corpus dans lesquels elles ont été observées. À partir de ce type de travaux, les contraintes préférentielles intervenant dans l'alternance dative ne peuvent pas être caractérisées de manière formelle. Par conséquent, il est difficile de différencier ce qui n'est qu'une tendance sur corpus de ce qui pourrait être considéré comme faisant partie de la connaissance des locuteurs.

1.3.3. Le travail de Bresnan *et al.* (2007)

Le travail sur corpus que nous allons présenter s'attache à répondre aux deuxième et troisième problèmes soulevés précédemment, grâce à l'utilisation d'outils statistiques plus sophistiqués. Ces outils permettent notamment de prendre en compte simultanément les variables et de tenter d'inférer des conclusions générales à partir de l'échantillon que constitue un corpus. Le premier problème évoqué, à savoir le nombre de contraintes préférentielles intervenant dans le phénomène, ne peut pas être complètement écarté.

Le travail de Bresnan *et al.* (2007) présente deux objectifs principaux. Le premier, d'ordre méthodologique, consiste à montrer que les données de corpus sont pertinentes pour comprendre le fonctionnement d'un phénomène langagier et qu'il existe des méthodes statistiques permettant de répondre aux principales critiques adressées aux données de corpus. Le deuxième objectif est d'ordre théorique et repose sur l'interprétation des données statistiques pour l'étude de l'alternance dative.

Cette étude s'appuie sur un corpus de 2360 occurrences d'alternance dative, extraites du *Switchboard* (Godfrey *et al.*, 1992). Ces données ont été annotées pour 14 variables :

1. la classe sémantique du verbe (*abstract*, *transfer of possession*, *future transfer of possession*, *prevention of possession*, *communication*) ;
2. l'accessibilité du destinataire (*donné* vs *non-donné*) ;
3. l'accessibilité du thème (*donné* vs *non-donné*) ;
4. la pronominalité du destinataire (*pronom* vs *non-pronom*) ;
5. la pronominalité du thème (*pronom* vs *non-pronom*) ;
6. le caractère défini du destinataire (*défini* vs *indéfini*) ;
7. le caractère défini du thème (*défini* vs *indéfini*) ;
8. le caractère animé du destinataire (*animé* vs *non-animé* [= *non-humain* et *non-animal*]) ;
9. la personne grammaticale du destinataire (*locale* vs *non-locale*) ;
10. le nombre du destinataire (*singulier* vs *pluriel*) ;

11. le nombre du thème (*singulier* vs *pluriel*) ;
12. le caractère concret du thème (*concret* vs *non-concret*) ;
13. le parallélisme structural dans le dialogue (présence du même type de structure dans le dialogue) ;
14. la longueur relative du destinataire et du thème (en nombre de mots, échelle logarithmique) ;

Les auteurs utilisent la régression logistique [29] (Agresti, 2007) pour modéliser l'alternance dative en fonction de l'ensemble de ces variables. L'idée de la régression logistique est de modéliser le comportement d'une variable binaire en fonction d'un ensemble de variables prédictrices. Plus exactement, elle permet d'estimer la probabilité qu'une des deux valeurs de la variable binaire soit réalisée, étant donné les variables prédictrices [30]. Dans le cas de l'alternance dative, la variable binaire est le type de constructions choisi (construction à double objet ou construction à SP datif). Les variables prédictrices, quant à elles, correspondent aux quatorze variables listées ci-dessus.

Bresnan *et al.* s'attachent à montrer qu'il est possible de modéliser un phénomène de langue sur corpus, malgré la corrélation entre les variables, les idiosyncrasies lexicales et l'échantillon limité que représente le corpus. En étudiant les propriétés du modèle statistique construit à partir de leurs données, les auteurs peuvent déterminer le rôle des différentes contraintes. En effet, chaque variable intégrée au modèle statistique se voit attribuer un coefficient interprétable [31] qui indique quelle construction la variable favorise et avec quelle force. De plus, la modélisation permet de s'assurer que le rôle joué par chacune des contraintes préférentielles n'est pas réductible à l'effet d'une ou de plusieurs autres variables. Pour synthétiser les résultats observés dans le modèle, les auteurs proposent le diagramme, reproduit en (31), qui présente l'alignement harmonique des propriétés linguistiques avec la position syntaxique.

(31)
donné ≺ non-donné
pronom ≺ non-pronom
animé ≺ non-animé
défini ≺ non-défini
plus court ≺ plus long
V **SN** SN
V **SN** SP

D'après ce schéma, le constituant possédant le plus de propriétés en gras a ten-

29. Pour une présentation détaillée, voir section 2.2.2, chapitre 2.

30. Bresnan *et al.* (2007) ne sont pas les premiers à utiliser la régression logistique pour l'alternance dative. Par exemple, Williams (1994) présente un travail, dans la lignée de Thompson (1990), qui consiste à utiliser la régression logistique sur un corpus de 168 phrases. Cependant, Bresnan *et al.* (2007) sont les premiers à utiliser un aussi grand échantillon et à étudier d'aussi près les propriétés du modèle pour en tirer des généralités sur le phénomène linguistique.

31. L'interprétation des coefficients attribués aux variables n'est pas triviale. Elle nécessite un certain nombre de précautions. Nous y reviendrons plus en détail dans la section 2.2 du chapitre 2.

dance à apparaître en position directement postverbale. Si ce constituant est le destinataire, c'est la construction à double objet qui est choisie ; si c'est le thème, c'est la construction à SP datif. Par rapport à ce schéma, l'intérêt du modèle statistique est de permettre non seulement de faire émerger les tendances, mais aussi de formaliser précisément le rôle de chaque variable et ainsi de prédire avec précision la probabilité d'une construction en tenant compte des contraintes préférentielles en présence. Le travail de Bresnan (2007b), que nous avons évoqué dans la section précédente, prolonge l'approche statistique sur corpus en testant les corrélations entre les observations faites à partir des données attestées et les jugements d'acceptabilité des locuteurs.

L'article de Bresnan *et al.* (2007) est une référence majeure pour la méthodologie en linguistique de corpus. Il est aussi essentiel au développement de la notion de contraintes préférentielles car il présente des moyens formels permettant de les identifier et de capter leur fonctionnement. Nous reviendrons en détail sur ces méthodes formelles dans le chapitre 2.

1.3.4. L'alternance dative dans différentes variétés de l'anglais

Deux articles que nous avons évoqués dans la section 1.1.2.3, ont prolongé le travail de Bresnan *et al.* (2007) et Bresnan (2007b), en comparant le phénomène de l'alternance dative dans plusieurs variétés de l'anglais : comparaison anglais américain / anglais néo-zélandais (Bresnan & Hay, 2008) ; comparaison anglais américain / anglais australien (Bresnan & Ford, 2010). Ces travaux mettent en évidence la finesse des observations et des généralisations que l'on peut obtenir sur corpus (Bresnan & Hay, 2008), ainsi qu'au moyen de la confrontation des données de corpus avec des données expérimentales (Bresnan & Ford, 2010).

Dans un premier temps, Bresnan & Hay (2008) montrent qu'une modélisation sur corpus peut rendre compte de variations subtiles entre deux variétés de l'anglais (américain et néo-zélandais). Le corpus utilisé comprend uniquement des occurrences du verbe *give* apparaissant avec la construction à double objet ou à SP datif. La partie néo-zélandaise contient 1127 occurrences extraites du corpus d'oral ONZE (*Origins of New Zealand English*). La partie américaine correspond au sous-corpus contenant le verbe *give* dans le données de Bresnan *et al.* (2007). Elle contient 1263 occurrences extraites du SWITCHBOARD et 404 occurrences du corpus arboré du *Wall-Street Journal*. Les 2794 phrases étudiées ont été annotées pour 9 variables :

1. la longueur du thème,
2. la longueur du destinataire,
3. la pronominalité du thème,
4. la pronominalité du destinataire,
5. l'accessibilité du thème,
6. l'accessibilité du destinataire,
7. le caractère animé du destinataire,
8. la classe sémantique du verbe (*transfer*, *communication*, *abstract*),
9. la variété d'anglais (néo-zélandais, américain parlé, américain écrit).

Chaque variable testée présente un effet significatif qui suit l'effet observé dans le modèle de Bresnan *et al.* (2007). Les auteurs observent que l'interaction de la variété d'anglais et du caractère animé du destinataire est statistiquement significative. Cela révèle que les destinataires non-animés ont plus tendance à être utilisés dans une construction à double objet en néo-zélandais qu'en américain. D'après cet article, l'une des différences entre anglais néo-zélandais et anglais américain repose sur le degré d'influence du caractère animé dans l'alternance dative. La modélisation utilisée permet de rendre compte de ce fait et de le quantifier.

Dans un deuxième temps, Bresnan & Ford (2010) ont étudié l'alternance dative en anglais australien et en anglais américain. Leur article s'appuie sur une version mise à jour du corpus de Bresnan *et al.* (2007) contenant 2349 occurrences d'alternance dative, avec 79% de constructions à double objet. Notons que les variables *accessibilité du thème* et *accessibilité du destinataire* ne font plus partie du modèle, car elles n'y participent pas significativement. Pour évaluer les différences en termes de préférences et donc d'effets des contraintes préférentielles selon les deux variétés, les auteurs s'appuient notamment sur deux études corrélationnelles. La première utilise un questionnaire de jugements d'acceptabilité. Bresnan & Ford observent que les locuteurs australiens sont plus sensibles à la contrainte de longueur que les américains. Plus précisément, si le destinataire est plus long que le thème, les Australiens favorisent la construction V SN_{theme} SP_{dest}, de façon plus marquée que les Américains.

La deuxième étude corrélationnelle a pour objectif de montrer que les contraintes préférentielles mises à jour sur corpus ne sont pas seulement à l'oeuvre dans la tâche métalinguistique que représente le jugement, mais également lors du traitement des phrases dans le processus de compréhension. Comme nous l'avons expliqué dans la section 1.1.1, Bresnan & Ford utilisent une tâche de lecture avec décision lexicale continue afin de recueillir des temps de réaction à la lecture de la préposition *to* dans la construction à SP datif. Les auteurs observent que la longueur du thème et son caractère défini ont un effet significatif sur le temps de décision lexicale. L'effet de la longueur est le suivant : plus le thème est long, plus le temps de réaction est élevé, car les sujets s'attendent de moins en moins à rencontrer une construction à SP datif. De plus, il y a une interaction significative entre les variables *longueur du thème* et *variété*. D'après cette interaction, les Américains sont plus sensibles à la longueur du thème : si le thème s'allonge, leur temps de réaction augmente beaucoup plus vite que celui des Australiens.

Bresnan & Ford (2010) concluent qu'il existe bien une différence d'effet de la contrainte de longueur dans les deux variétés. Elles interprètent les résultats des deux études corrélationnelles en émettant l'hypothèse que les locuteurs australiens ont une préférence plus marquée pour la construction à SP datif que les Américains. Ainsi, l'effet d'une contrainte favorisant l'ordre SN_{theme} SP_{dest}, comme *destinataire plus long que thème*, est plus fort chez les Australiens, et, inversement, l'effet d'une contrainte favorisant l'ordre SN_{dest} SN_{theme}, comme *thème long et complexe*, est moins fort chez les Australiens que chez les Américains.

L'exemple des travaux sur l'alternance dative montre qu'une approche en termes de contraintes préférentielles permet de rendre compte du phénomène de façon sa-

tisfaisante. Du point de vue méthodologique, ces travaux montrent comment il est possible de tirer des généralités à partir de corpus et que les études corrélationnelles fournissent un outil complémentaire permettant de s'assurer que les phénomènes observés sont en correspondance avec une forme de savoir langagier.

1.4. Quel objet pour la syntaxe ?

En intégrant les contraintes préférentielles dans le champ de la syntaxe, on élargit l'objet de cette dernière à des dimensions non-catégoriques.

Nous considérons que l'architecture fondamentale du langage se définit en termes absolus. Les contraintes catégoriques définissent les structures de base de la langue, au sein desquelles les contraintes préférentielles prennent place. Par exemple, dans le syntagme nominal français, la position du déterminant et des dépendants du nom non-adjectivaux est définie catégoriquement : le déterminant précède obligatoirement le nom et les dépendants non-adjectivaux lui succèdent. La position de l'adjectif par rapport au nom est, quant à elle, définie selon des contraintes préférentielles, favorisant plus ou moins fortement l'antéposition ou la postposition [32].

Dans cette perspective, l'objet d'étude de la syntaxe ne concerne pas seulement une compétence définie en termes de grammaticalité, mais aussi une connaissance langagière mettant en jeu des préférences. Cette compétence langagière rejoint la notion de « *conception du savoir partagé* » [33] que Wasow (2009) définit de la façon suivante : « *un corps commun de savoir auquel nous faisons appel pour différents usages du langage, parmi lesquels comprendre ce que les autres disent, formuler nos propres énoncés, émettre des jugements sur des phrases exemples, lire, écrire, traduire, paraphraser, faire des jeux de mots etc* » [34]. Cette vision ne s'oppose pas à celle de la grammaire générative, mais elle élargit l'objet d'étude de la syntaxe et met en avant le fait que les connaissances linguistiques sont avant tout des connaissances permettant de produire et de comprendre des énoncés.

La compétence langagière des locuteurs repose sur des aspects non-catégoriques qui sont, en partie, déterminés par la fréquence d'occurrences des constructions. Nous émettons l'hypothèse selon laquelle la compétence est façonnée, en partie, par l'usage. Notons que cette vision est en accord avec la linguistique diachronique (Beckner & Bybee, 2009; Bybee, 2006, 2009, parmi d'autres), selon laquelle la fréquence d'usage est un facteur essentiel de grammaticalisation et d'évolution de la langue. Bybee (2009) résume les effets de la fréquence comme suit : (1) les mots et les séquences très fréquents subissent de rapides réductions phonétiques ; (2) ils sont renforcés dans leur structure morpho-syntaxique et résistent donc à des règles plus générales ; (3) ils peuvent perdre leur structure interne et devenir autonomes par rapport à des formes

32. Les chapitres 3 et 4 seront consacrés à cette question.

33. Traduction de « *"shared knowledge" conception* ».

34. « *…a common body of knowledge that we draw on for distinct uses of language, including understanding what others say, formulating our own utterances, making judgments regarding example sentences, reading, writing, translating, paraphrasing, making puns, etc.* »

étymologiquement liées. Ces effets de fréquence sont un des moteurs essentiels des changements linguistiques.

Enfin, en affirmant que les contraintes préférentielles sont des contraintes linguistiques ayant leur place dans la compétence langagière des locuteurs, nous remettons en cause l'hégémonie de la notion de grammaticalité en syntaxe. Comme le dit Manning (2003), la notion de grammaticalité catégorique est trop puissante par rapport à la réalité du langage : « *Les théories linguistiques font trop d'affirmations. Elles placent une frontière de grammaticalité catégorique dure, là où il y a en réalité une limite floue, déterminée par de nombreuses contraintes contradictoires et par des questions de conformisme vs créativité humaine* »[35]. Nous considérons que la notion d'acceptabilité entre en jeu dans une série de phénomènes syntaxiques qui ont longtemps été envisagés en termes catégoriques. Nous proposons, dans la lignée des travaux effectués sur l'alternance dative, de les explorer en nous fondant sur des méthodes d'analyse statistique appliquées à des données expérimentales et de corpus.

35. « *Categorical linguistic theories claim too much. They place a hard categorical boundary of grammaticality where really there is a fuzzy edge, determined by many conflicting constraints and issues of conventionality vs human creativity* » (Manning, 2003, p. 297)

Chapitre 2

Méthodes et Outils

Sommaire

L'objet de ce chapitre est d'introduire des méthodes d'analyse de données en vue de les appliquer au domaine de la syntaxe. Ces méthodes ont déjà été utilisées dans d'autres sciences humaines (psychologie, sociologie, économie), comme le prouve, par exemple, l'ouvrage de Howell (1998). En revanche, leur application à la syntaxe ne s'est développée que très récemment.

Pour mener une étude sur les contraintes préférentielles, il faut des données. Comme nous l'avons montré dans le chapitre précédent, les données issues de l'introspection et des jugements de grammaticalité ne sont pas suffisantes pour étudier des phénomènes impliquant des contraintes préférentielles. Deux autres sources de données peuvent être envisagées : les données de corpus et les données expérimentales. Nous appuyons nos recherches principalement sur des données de corpus. Toutefois, nous utiliserons également des données expérimentales dans le but de pallier les difficultés inhérentes au recueil et à l'analyse des données de corpus.

Notre démarche consiste à travailler sur des données de corpus. Cependant, cette démarche pose le problème de la représentativité des données analysées. En effet, idéalement, l'échantillon étudié devrait être représentatif de la langue, ce qui nous permettrait de pouvoir tirer des généralités sur le système de la langue à partir de l'analyse statistique de cet échantillon. L'autre enjeu de la démarche utilisée est la généralisation : comment passer des observations sur corpus à des propriétés du système de la langue ? Traditionnellement, en syntaxe, le processus de généralisation consiste à proposer des hypothèses sur le système de la langue permettant de rendre compte des faits observés (cf. section 1.2.1 du chapitre précédent). Les statisticiens, pour leur part, cherchent à inférer des propriétés d'une population à partir d'un échantillon, c'est-à-dire à avoir une connaissance générale de la population à partir de données sur une partie de la population. Par abus de langage, nous utiliserons le terme 'généralisation', là où les statisticiens parleraient d'inférence. Nous envisagerons donc la généralisation vers le système de la langue grâce aux outils des statisticiens.

Bien que les méthodes d'analyse de données ne soient pas exemptes de limitations, notamment en ce qui concerne la représentativité et la généralisation, nous estimons que ce type d'approche fournit une méthodologie qui permet de compléter les approches classiques en syntaxe (grammaticalité et introspection).

Ce chapitre se compose de trois sections. La première aura pour objet la notion de corpus ainsi que les données qui en sont extraites. Nous aborderons le problème de la représentativité des corpus, ainsi que celui de la généralisation. Nous présenterons également les corpus du français d'où nous avons extrait les données étudiées dans cette thèse. La deuxième section sera consacrée aux outils statistiques permettant l'analyse des données de corpus. Nous y introduirons l'outil de modélisation que nous utiliserons dans les chapitres suivants, à savoir la régression logistique. Pour cela, nous exposerons d'abord les modèles de régression linéaire ainsi que la méthodologie qui leur est associée. Nous présenterons ensuite les modèles logistiques, en nous appuyant sur les méthodes et les concepts développés pour les modèles linéaires. Enfin, dans la troisième section, nous aborderons la question de l'expérimentation en linguistique et

celle des données issues d'expériences psycholinguistiques, en présentant deux types d'expérience que nous utiliserons.

2.1. Obtenir les données : le corpus

La première source de données que nous utilisons est le corpus. Ce dernier est un objet fabriqué à partir de textes ou de discours produits dans des conditions "naturelles" ou "écologiques", à savoir, des conditions où les locuteurs parlent ou écrivent sans savoir que la séquence produite sera analysée à des fins linguistiques. Les données "naturelles" s'opposent donc aux données "construites" par leurs conditions de production. De plus, les données "construites" sont généralement étudiées en fonction de jugements que des locuteurs ont à leur égard, tandis que les données de corpus, telles qu'elles sont envisagées dans notre travail, représentent un échantillon de langue analysé sur la base des fréquences observées [1]. En tant qu'objet fabriqué par les linguistes, le corpus n'est pas un objet vierge de toute théorie. Par exemple, le découpage en mots imposé par les conventions d'écriture du français est une représentation orientée de la langue : par cette convention, les clitiques sont considérés comme des mots, alors que certains travaux linguistiques ont montré qu'ils pourraient être analysés comme des préfixes du verbe (Miller, 1992; Miller & Sag, 1997). Cet exemple très simple révèle que le comptage de mots sur un corpus écrit "brut" ne représente pas une donnée brute, mais bien une donnée construite, reposant sur des hypothèses linguistiques implicites. Cette observation est encore plus manifeste lorsque l'on utilise des corpus oraux transcrits. La construction de ces corpus engendre un nombre très important de choix qui peuvent avoir des conséquences sur les analyses des données (Dister & Simon, 2008; Mondada, 2000). Notre travail repose sur l'utilisation d'annotations en parties du discours et en structures syntaxiques. Cela implique que les données utilisées s'appuient sur des descriptions linguistiques existantes. Les études que nous développons reposent donc sur l'hypothèse que ces descriptions sont justifiées et adéquates.

2.1.1. La représentativité du corpus

Les études quantitatives posent le problème de la représentativité des données de corpus. Les données formant le corpus représentent un échantillon par rapport à la population que constitue la langue. Autrement dit, un corpus est un tout petit extrait de phrases produites sur l'ensemble des phrases produites dans une langue donnée. Le principe général de l'analyse statistique consiste à induire, à partir de l'échantillon, des propriétés quantitatives de la population. Par exemple, dans les sondages d'opinion, le sondeur cherche à avoir une image des opinions de la popula-

1. Les données "naturelles" et les données "construites" constituent deux outils à la disposition du linguiste. Ce sont des sources de données complémentaires : les données de corpus sont plus adaptées à un travail sur des préférences, alors qu'une étude sur un phénomène syntaxique rare sera plus facile à aborder avec des données construites.

tion à partir d'un échantillon de la population. De la même façon, en linguistique, on cherche à inférer des propriétés quantitatives valables pour la langue en se fondant sur les fréquences observées dans le corpus dont on dispose. Cependant, les fréquences observées dans l'échantillon ne sont pas toutes représentatives des fréquences de la population. Certaines sont des artefacts dus à l'échantillonnage, c'est-à-dire au hasard. L'objectif idéal de l'analyse de données est donc d'inférer des propriétés quantitatives de la population à partir de l'échantillon, tout en écartant les propriétés dues au hasard.

Pour cela, il faut s'appuyer sur une méthode d'échantillonnage qui assure la qualité, la représentativité de l'échantillon. Dans le cas des sondages d'opinion, les instituts de sondage procèdent généralement à ce que l'on appelle un échantillonnage empirique par quotas. Selon cette méthode, l'échantillon retenu doit avoir la même composition que la population par rapport à une ou plusieurs caractéristiques (sexe, âge, catégories socio-professionnelles…). En ce qui concerne la linguistique, la méthode par quotas n'est pas envisageable dans la mesure où les caractéristiques de la population entière ne sont pas connues. La méthode d'échantillonnage qui semble la plus adaptée est l'échantillonnage aléatoire stratifié.

Pour expliciter cette méthode et son application à la linguistique de corpus, nous reprenons la métaphore de la bibliothèque (*library metaphor*) de Evert (2006) : « *Imaginons une immense bibliothèque qui représente l'intégralité d'une langue, ou d'une sous-langue, comme objet d'étude. Chaque livre de la bibliothèque correspond à un fragment de la langue – certains grands, d'autres petits – qui peut être utilisé comme un corpus linguistique. Sélectionner ou compiler un corpus revient alors à choisir un livre au hasard sur l'une des étagères…* » [2]. D'après cette métaphore, un ensemble d'articles du journal *Le Monde* ou la transcription de plusieurs heures de radio constituent un fragment de langue aléatoirement choisi. Il s'agit donc d'un échantillon de langue à partir duquel on peut faire des observations quantitatives. Cependant, cette vision est un peu simpliste car elle suppose que la langue est un objet homogène qui n'est pas composé de sous-ensembles. Or, il existe différents genres et sous-genres qui présentent des particularités linguistiques (par exemple, genre journalistique, littéraire, oral spontané…). Si l'on reprend la métaphore de la bibliothèque, il existe des sections dans lesquelles sont classés les livres. De plus, chaque livre de la bibliothèque présente un style différent selon l'auteur, un contenu lexical différent selon le thème abordé. Ainsi, choisir un seul livre dans la bibliothèque implique un double risque de biais : le biais du genre du livre et le biais du livre lui-même. Evert (2006) poursuit sa métaphore en expliquant que pour constituer un corpus non biaisé, il faudrait extraire aléatoirement des phrases, ou des ensembles de phrases, dans des livres eux-même choisis aléatoirement dans la bibliothèque. Pour que le corpus soit représentatif, il faudrait également que chaque section de la bibliothèque soit représentée, afin que chaque genre apparaisse dans l'échantillon étudié. Les corpus tels que le

2. « *imagine a gigantic library that represents the entirety of a language or sublanguage as the object of study. Each book in this library corresponds to a fragment of the language – some large, some small – that could be used as a linguistic corpus. Compiling a corpus, thus, amounts to picking a book at random from one of the shelves.* » (Evert, 2006, p. 178)

British National Corpus ou le *Brown Corpus* peuvent être considérés comme des corpus non biaisés et représentatifs car ils contiennent une grande variété d'échantillons linguistiques tirés d'un grand nombre de documents différents (Evert, 2006, p. 183). La variété des documents et des genres représente un grand pas vers la constitution d'un corpus représentatif. Néanmoins, il existe une limite difficile à surmonter : pour obtenir un corpus réellement représentatif de la population, il faudrait que le nombre de phrases extraites de chaque section de la bibliothèque soit proportionnel à la taille de chaque section. C'est seulement à cette condition que la méthode d'échantillonnage aléatoire stratifiée serait respectée. Cependant, cela représente une difficulté actuellement indépassable car nous ne connaissons pas la taille des sections de la bibliothèque. Comme le précise Evert (2006), la méthode adoptée est d'émettre des hypothèses sur la taille de ces sections et donc sur l'importance relative de chaque genre. Ces hypothèses sont largement influencées par les moyens dont les chercheurs disposent.

Dans le cas de l'étude du français, nous ne disposons pas d'un corpus tel que le *British National Corpus*. Les corpus utilisés ont d'abord été choisis en fonction de leur accessibilité et de la possibilité de les utiliser pour les phénomènes qui nous intéressent. Notons que pour les deux phénomènes que nous allons étudier, la position de l'adjectif épithète et l'ordre des compléments postverbaux, nous n'utiliserons pas les mêmes corpus. Nous ne pouvons pas considérer que les corpus utilisés dans ces deux études de cas sont représentatifs de la langue française. Cependant, nous avons tenté d'apporter des éléments pour aider à la généralisation. Dans le cas des compléments postverbaux, nous avons étendu au maximum les sources de données afin de rendre compte, le plus possible, de la variété des genres existants. En ce qui concerne les adjectifs, l'étude sur corpus reste limitée à un seul corpus journalistique, mais les résultats ont été confrontés à la connaissance des locuteurs à travers une étude corrélationnelle [3].

2.1.2. Qu'est-ce que l'on compte ?

Pour parler de la représentativité des corpus, nous avons utilisé comme unité d'échantillonnage la phrase. Or, l'unité d'échantillonnage dépend du phénomène auquel on s'intéresse (Baroni & Evert, 2008, p 18). Les échantillons étudiés dépendent donc des phénomènes étudiés.

Dans les faits, nous partons d'un échantillon de langue déjà disponible sous la forme d'un corpus. C'est à partir de ce premier échantillon que l'on crée l'échantillon qui correspond à la population qui nous intéresse. Dans le cadre de cette thèse, nous étudions deux phénomènes différents, ce qui signifie que nous visons deux populations différentes. Dans le cas de la position des adjectifs épithètes, l'unité étudiée est le syntagme nominal contenant une tête nominale et au moins un adjectif épithète. Dans ce cas, la population visée est l'ensemble des syntagmes nominaux ayant une tête

3. La distinction entre *expérience* et *étude corrélationnelle* a été exposée dans la note 2, de la section 1.1.1 du chapitre 1. Nous reviendrons en détail sur l'idée d'étude corrélationnelle à la section 2.3.

nominale et au moins un adjectif épithète. En ce qui concerne l'ordre des compléments postverbaux, l'unité qui nous intéresse est une phrase contenant un verbe et ses deux compléments en position postverbale. La population visée est donc l'ensemble des phrases présentant ces caractéristiques. À partir des échantillons composés des unités décrites, nous étudierons les facteurs qui ont un effet significatif sur la position de l'adjectif et sur l'ordre des compléments, en tentant de généraliser à l'ensemble des populations visées.

Evert (2006) soulève un autre point concernant l'utilisation des méthodes statistiques : l'analyse statistique des données linguistiques repose sur une vision extensionnelle de la langue. Le principe sur lequel s'appuie l'analyse statistique est que, à partir des données d'un échantillon, on généralise à l'ensemble de la population. Cela signifie qu'à partir d'un corpus de phrases, on généralise à l'ensemble des phrases de la langue. La démarche statistique amène donc à tirer des conclusions sur l'ensemble des unités étudiées de la langue, autrement dit sur le phénomène d'un point de vue extensionnel. Le passage de cette vision extensionnelle à une vision intensionnelle, où la langue est considérée comme un système, revient au chercheur. L'analyse statistique apporte donc des éléments sur la significativité de la fréquence de certaines unités dans la langue et il incombe au chercheur d'interpréter ces éléments dans le but de défendre un point de vue sur le système de la langue.

2.1.3. Corpus utilisés

Dans cette partie, nous présentons les corpus de français exploités dans cette thèse. Dans les parties consacrées aux deux grands phénomènes que nous étudions, nous reviendrons plus en détail sur l'utilisation de ces corpus et sur les méthodes de collecte des données spécifiques à chaque sujet. Les données dont nous nous servons sont extraites de ressources linguistiques existantes. Plus précisément, nous avons utilisé deux corpus journalistiques (*French Treebank* et Est-Républicain), un corpus de radio transcrite (ESTER) et un corpus d'oral spontané (CORAL-ROM).

French Treebank (FTB) Le corpus FTB est la seule "banque d'arbres" existante pour le français. Elle est le résultat d'un projet, mené à l'université Paris Diderot sous la direction d'Anne Abeillé, qui avait pour objectif l'annotation supervisée d'articles du journal *Le Monde* datant de 1989 à 1993. Ce corpus, disponible depuis 2003, est composé de 32 000 phrases qui sont constituées de 870 000 tokens représentant 37 000 lemmes (Abeillé *et al.*, 2003). Ce corpus est annoté pour les parties du discours, la flexion, les composés, les lemmes et la constituance.

Nous utilisons la sous-partie du FTB annotée en fonctions grammaticales (Abeillé & Barrier, 2004). Ce sous-corpus compte 12 351 phrases, 24 898 lemmes et 385 458 tokens. Nous en avons utilisé une version pour laquelle l'annotation des mots composés a été neutralisée. Ces derniers sont considérés comme des formes uniques non structurées.

Est-Républicain (ER) Le corpus ER est constitué du texte intégral de deux années du journal régional de l'Est-Républicain. Il se compose de 148 millions de tokens répartis en 662 000 lemmes. Ce corpus est disponible sur le site du CNRTL [4].

La version du corpus que nous utilisons est une version enrichie de façon automatique et librement disponible (Seddah *et al.*, 2012). Elle comporte une annotation en lemmes et en parties du discours qui a été réalisée grâce au modèle MORFETTE décrit par Chrupała *et al.* (2008) et adapté au français par Seddah *et al.* (2010).

ESTER Le corpus ESTER est issu du projet Evalutation des Systèmes de Transcription d'Émissions Radiophoniques. Distribué par ELRA [5], il se compose d'une partie sonore et d'une partie transcrite de soixante heures de radio. Les données radiophoniques datent de 2002 à 2005 et sont issues de France Inter, France Info, France Culture, RFI, Radio Classique et de la Radio Télévision Marocaine. Nous utilisons la partie transcrite qui contient 707 553 mots et qui est annotée pour les parties du discours, les lemmes et la flexion.

CORAL-ROM Le corpus CORAL-ROM est un corpus multilingue de parole spontanée pour quatre langues romanes (français, italien, portugais et espagnol), distribué par ELRA [6]. Il compte environ 300 000 mots par langue et contient des conversations et des monologues produits dans des contextes formels et informels. Il est annoté en parties du discours et en lemmes.

Les données utilisées sont issues de corpus écrits journalistiques et de corpus oraux de deux types : oral surveillé (ESTER) et oral spontané (CORAL-ROM). Notons que malgré un genre commun, les corpus FTB et ER présentent un style et un contenu très différents. Le premier fait apparaître un style soutenu et aborde des sujets économiques et politiques d'ordre national et international, tandis que le second utilise un style moins soutenu et traite d'actualités locales (sport, fait divers, culture...). Ces deux corpus ne sont donc pas redondants et permettent, au contraire, d'observer une variété thématique et lexicale. Dans le cas de ESTER, les thèmes abordés sont proches de ceux abordés dans FTB (actualités nationales et internationales), mais les deux se distinguent par leur mode de transmission (oral vs. écrit). Enfin, dans les deux corpus d'oral spontané, des thèmes très variés sont abordés. Les situations de communication et les types d'interaction sont également largement diversifiés, ce qui permet une grande variété lexicale et syntaxique.

Une fois les données recueillies et décrites, se pose le problème de leur analyse qui doit permettre le passage de la simple observation de faits dans l'échantillon vers la généralisation à une population plus large : une sous-langue ou la langue elle-même.

4. Site du corpus Est-Républicain : http://www.cnrtl.fr/corpus/estrepublicain/.
5. http://catalog.elra.info/product_info.php?products_id=999.
6. http://catalog.elra.info/product_info.php?products_id=757

2.2. Analyses de données

L'objectif de cette partie est à la fois d'expliciter les outils d'analyse statistique que nous utiliserons, et de présenter la pertinence et l'utilité de ce type de méthode appliquée à la linguistique, en insistant notamment sur la possibilité de conclure au-delà de l'échantillon. Il ne s'agit pas de faire une introduction aux statistiques et aux méthodes de régression, mais d'introduire et de rendre accessibles les notions essentielles à la formalisation et à la généralisation à partir de données de corpus et de données expérimentales.

L'analyse de données vise à déterminer des caractéristiques générales de la population langue à partir de l'échantillon dont on dispose. L'idée de chercher la généralisation à partir d'un échantillon est très répandue dans les autres sciences humaines mais a été longtemps bannie de la linguistique. Comme le note Gries (2009) : « *il peut apparaître surprenant que les méthodes statistiques ne soient pas très répandues en linguistique. Cela est d'autant plus surprenant que de telles méthodes sont très répandues dans des disciplines traitant de sujets aussi complexes, telles que la psychologie, la sociologie, l'économie. À un certain degré, cette situation est probablement due à la façon dont la linguistique a évolué durant les décennies passées, mais heureusement, cela est en train de changer maintenant* » [7]. Il semble que la linguistique pourrait largement bénéficier de l'apport de méthodes quantitatives, car elles constituent un outil supplémentaire mis à la disposition du linguiste pour lui permettre de décrire et de chercher la généralisation. Certaines branches de la linguistique font déjà usage des méthodes statistiques : la psycholinguistique et la sociolinguistique, sous l'influence respective des méthodes en psychologie et en sociologie. Nous citerons d'ailleurs quelques exemples psycholinguistiques simples afin d'illustrer la régression linéaire multiple et la régression à effets mixtes.

Logiciel de statistique et données utilisées L'ensemble des analyses statistiques et des graphiques proposé ici a été exécuté avec R (R Development Core Team, 2011). Il s'agit d'un logiciel gratuit et librement accessible. Il a été conçu pour le traitement des données, pour l'analyse statistique ainsi que pour la réalisation de graphiques. L'environnement R ainsi que des extensions librement téléchargeables, offrent, en plus des outils statistiques, des jeux de données. Dans le cadre de ce chapitre, nous illustrerons notre propos en utilisant quatre de ces jeux de données. Le premier est disponible dans la version de base de R et les trois suivants viennent avec l'extension `languageR` créée par Baayen (2008) et largement dédiée à l'étude du langage. Les quatre jeux de données sont :

7. « *…it may appear surprising that statistical methods are not that widespread in linguistics. This is all the more surprising because such methods are very widespread in disciplines with similarly complex topics such as psychology, sociology, economics. To some degree, this situation is probably due to how linguistics has evolved over the past decades, but fortunately it is changing now.* » (Gries, 2009, p. 4).

- *cars* : ces données regroupent 50 mesures de vitesse de voiture associées à une distance d'arrêt.
- *lexicalMeasures* : cet ensemble de données contient, pour 2233 mots monomorphémiques de l'anglais, des mesures lexicales distributionnelles telles que la longueur, la fréquence, le nombre de voisins orthographiques…
- *lexdec* : ce jeu de données rassemble les temps de décision lexicale de 21 sujets pour 79 noms concrets anglais, ainsi que des informations relatives aux sujets (langue maternelle, sexe) et aux mots testés (fréquence, longueur, classe sémantique…).
- *dative* : ces données, relatives à l'alternance dative en anglais, comprennent le type de réalisation du destinataire (SN ou SP) pour 3263 syntagmes verbaux recueillis dans le corpus *Switchboard* (Godfrey *et al.*, 1992) et dans le corpus *Treebank Wall Street Journal*. Pour chaque réalisation, la table de données contient des informations linguistiques concernant le verbe, le destinataire et le thème.

Nous avons rédigé les parties relatives à la régression linéaire simple et multiple à partir de deux ouvrages : Howell (1998) et Judd *et al.* (2010). En ce qui concerne les parties plus spécifiques au langage, notamment les données *lexdec* et *dative*, Baayen (2008) a été notre principale source d'inspiration. Enfin, les parties portant sur les modèles de régression un peu plus sophistiqués ont bénéficié des apports de Agresti (2007) pour la régression logistique et de Gelman & Hill (2007) et Pinheiro & Bates (2000) pour la régression à effets mixtes.

Afin de présenter l'analyse statistique des données, nous allons procéder en deux parties. Dans un premier temps, nous présenterons la régression linéaire qui est le cas le plus simple de régression. Après avoir expliqué en quoi consiste cet outil de modélisation et comment il est construit, nous introduirons les procédures méthodologiques permettant d'évaluer la qualité de la régression, de choisir les variables intervenant dans la régression ainsi que d'interpréter le modèle de régression obtenu. Nous introduirons également un modèle un peu plus complexe : la régression à effets mixtes, qui permet de mieux respecter la façon dont sont structurées les données. Dans un deuxième temps, nous présenterons la régression logistique, méthode de modélisation qui est au centre de cette thèse. Nous adopterons la même organisation que pour la régression linéaire : après avoir présenté l'outil de modélisation et sa mise en oeuvre, nous exposerons les procédures méthodologiques qui permettent de garantir la qualité de la modélisation.

Une partie des éléments présentés ici repose sur deux notions essentielles en statistique : la distribution d'échantillonnage et le test d'hypothèse. Pour une introduction à ces notions, le lecteur pourra se reporter, par exemple, au chapitre 4 de Howell (1998) ainsi qu'au chapitre 3 de Vasishth & Broe (2011).

Vitesse	Distance
6.4	0.6
6.4	3.0
11.3	1.2
11.3	6.7
12.9	4.9
14.5	3.0
…	…

TABLE 2.1.: Extrait de la table de données *cars*.

2.2.1. Régression linéaire

La régression linéaire est une méthode statistique qui consiste à modéliser le comportement d'une variable à partir d'une ou plusieurs autres variables, en s'appuyant sur l'hypothèse simplificatrice selon laquelle la relation qu'entretiennent ces variables est linéaire. L'intérêt de cette méthode est de permettre de prédire, pour toute nouvelle valeur des variables prédictrices, une valeur pour la variable à prédire. Dans cette partie, nous expliquons comment est construit un modèle de régression linéaire à partir des données et comment ce modèle permet de généraliser au-delà des valeurs de l'échantillon. Nous présenterons trois régressions par ordre de complexité : la régression simple, la régression multiple et la régression à effets mixtes [8].

2.2.1.1. Régression simple

Dans le cas le plus simple de la régression, nous disposons de deux ensembles d'observations représentés sous la forme de deux variables : l'une est la variable à prédire, notée y, l'autre la variable prédictrice, notée x. Le principe de la régression linéaire est de travailler avec l'hypothèse selon laquelle il existe une dépendance linéaire entre la donnée prédictrice x et la donnée à prédire y, de la forme :

$$y = \alpha + \beta x \tag{2.1}$$

Les exemples de telles dépendances abondent dans le monde de la physique. Prenons l'exemple simple de la dépendance entre la vitesse et la distance. Théoriquement, cette dépendance est linéaire. Afin d'illustrer cette idée, observons les données *cars* qui présentent 50 mesures de vitesse de voiture associées à une distance d'arrêt [9]. Un extrait de la table de données est présenté dans la table 2.1.

8. Le rapport entre ces trois régressions n'est pas symétrique : la régression simple est un cas particulier de régression multiple, alors que la régression linéaire à effets mixtes est un autre type de modélisation qui implique l'utilisation d'autres méthodes d'estimation et d'évaluation, comme nous le verrons dans la suite de cette section.

9. Les données originales sont en miles et en pieds, nous les avons converties en km/h et en mètres. Les mesures datent de 1920 !

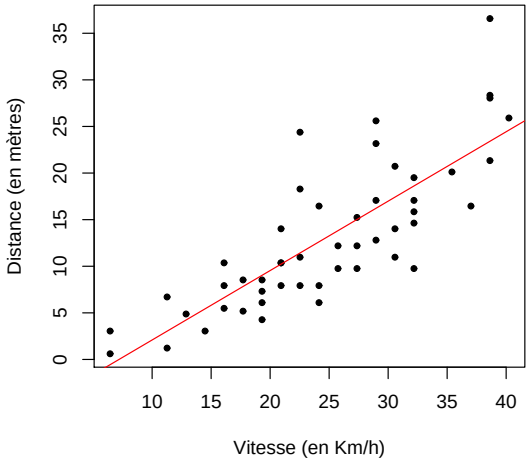

FIGURE 2.1.: Diagramme de dispersion des données de *cars*, avec la droite de régression $y = \alpha x + \beta$ où $\alpha = -5.35785$ et $\beta = 0.7447$

La figure 2.1 est un diagramme de dispersion de ces données avec la vitesse en abscisse et la distance en ordonnée. La droite rouge est la droite de régression : c'est la droite qui résume le mieux la relation entre vitesse et distance d'arrêt dans ces données. Une droite se définit par une équation de la forme de celle présentée en 2.1, où β, la pente, représente l'inclinaison de la droite, et α, l'intercept, représente la valeur de y quand x est égal à 0. La droite de la figure 2.1 constitue une représentation graphique de l'estimation de la dépendance linéaire théorique qui existe entre la variable à prédire, la distance, et la variable prédictrice, la vitesse. Nous constatons que les données observées présentent une variabilité par rapport à la dépendance théorique. Le but de la régression est de donner une estimation de la dépendance théorique à partir des données expérimentales, en dépit de la variabilité de ces données[10]. À cette fin, nous utilisons ce qu'on appelle les résidus. Le résidu représente la distance entre un point observé et le point correspondant sur la droite de régression. Soit $\hat{y}_i$ la distance correspondant à une vitesse donnée x selon la droite de régression et soit y_i la distance effectivement observée pour la vitesse x, le résidu correspond à la différence entre la valeur prédite $\hat{y}_i$ et la valeur observée y_i. Cela se note de la façon suivante, $\epsilon_i = y_i - \hat{y}_i$. Graphiquement, le résidu correspond à la distance entre une observation et sa projection verticale sur la ligne de régression. Pour les données *cars*, chaque résidu est représenté par une ligne bleue sur la figure 2.2. Ces résidus s'interprètent comme des erreurs de prédiction par rapport à la droite de régression. Trouver la droite de régression revient à calculer les valeurs de α et de β de l'équation 2.1, pour

10. De manière générale, une mesure expérimentale est entachée d'imprécision. Dans le cas de l'exemple de *cars*, les instruments de mesure ont pu produire des distances ou des vitesses inexactes.

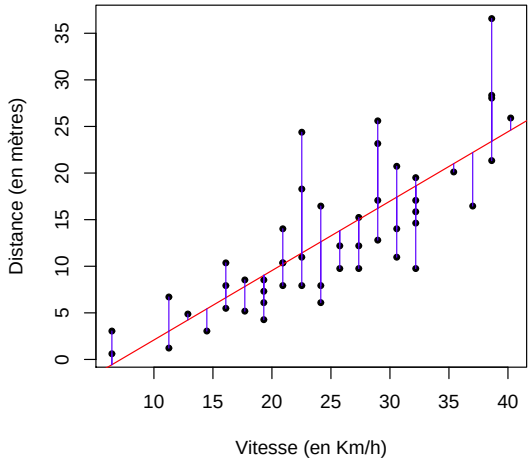

FIGURE 2.2.: Diagramme de dispersion des données de *cars*, avec la droite de régression en rouge et les résidus de chaque observation en bleu.

lesquelles l'erreur est minimale. La droite de régression linéaire est définie comme étant celle qui minimise l'erreur quadratique, c'est-à-dire la somme des carrés des résidus. La somme des carrés des résidus SC s'exprime comme une fonction prenant α et β comme arguments :

$$SC(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 \tag{2.2}$$

où n représente le nombre de données observées et où x_i et y_i sont des données fixes. La méthode qui consiste à chercher les valeurs de α et de β pour lesquelles la quantité de $SC(\alpha, \beta)$ est minimale s'appelle la méthode des moindres carrés. En suivant l'hypothèse selon laquelle les résidus sont distribués normalement [11] autour de la droite de régression, on obtient des équations nous permettant d'estimer α et β tels que la valeur de SC soit minimum [12]. Le modèle linéaire construit à partir des données d'observation se résume par la fonction suivante :

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{2.3}$$

11. Pour une définition du concept de distribution normale, se reporter au chapitre 3 de Howell (1998).

12. Calcul de α et β, avec $\bar{x}$ et $\bar{y}$ les moyennes des mesures x_i et y_i :

$$\alpha = \bar{y} - \beta \bar{x}$$
$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

où ϵ_i représente l'imprécision sur la mesure y_i. La notation utilisée pour reporter les résidus d'un modèle est $\epsilon \sim N(\mu, \sigma^2)$. Elle dénote le fait que les résidus se distribuent selon une loi normale de moyenne μ et de variance σ^2 [13]. Par exemple, pour les données *cars*, le modèle a la forme suivante : $\texttt{Distance} = -5.36 + 0.74 \times \texttt{Vitesse}, \epsilon \sim N(0, 4.687)$.

Pour récapituler, la régression linéaire simple permet de résumer un nuage de points en utilisant la droite la mieux ajustée aux données. La construction du modèle revient à trouver la droite de régression pour laquelle l'erreur est minimale.

2.2.1.1.1. Évaluation du modèle Une fois que le modèle de régression a été construit à partir des données de l'échantillon, il faut l'évaluer. Nous présentons deux aspects de l'évaluation. Premièrement, nous exposons une méthode permettant de connaître la validité du modèle construit sur l'échantillon par rapport à la population visée. Pour cela, nous faisons un test d'hypothèse sur le coefficient β. Deuxièmement, le modèle est évalué en fonction de ses capacités à prédire des valeurs proches des valeurs effectivement observées. Plus les valeurs prédites sont proches des valeurs observées, plus le modèle est considéré comme bon.

Test d'hypothèse sur le coefficient β L'objectif de cette sous-section est d'évaluer si le coefficient de régression estimé dans l'échantillon est bien différent de 0 dans la population à partir de laquelle l'échantillon a été extrait. Autrement dit, le but est de savoir s'il existe une relation linéaire entre les deux variables dans la population et si la relation observée n'est pas un simple artéfact dû à l'échantillonnage.

Le coefficient de régression β représente la relation linéaire unissant la variable à prédire et la variable prédictrice. Si ce coefficient est égal à 0, il n'existe pas de relation linéaire entre les deux variables : elles sont indépendantes. Graphiquement, si $\beta = 0$, la droite est horizontale. Cela signifie que pour n'importe quelle valeur de la variable prédictrice x, la droite de régression donne la même valeur à y (et cette valeur est égale à α, l'intercept). Ce cas de figure est illustré par le graphique 2.3, où la droite de régression a pour équation : $y = 3.5 + 0 \times x$.

Lorsque nous construisons un modèle linéaire, le coefficient β est estimé pour les données étudiées, c'est-à-dire pour un échantillon de la population. En calculant ce coefficient sur l'échantillon, on a pour objectif l'estimation du coefficient sur la population entière. De cette façon, nous espérons décrire la relation qui unit la variable prédictrice et la variable à prédire dans la population entière et ainsi pouvoir donner une valeur à la variable à prédire pour n'importe quelle nouvelle valeur de la variable prédictrice. Cependant, la valeur de β n'est pas et ne sera jamais la valeur exacte du coefficient β dans la population totale. Ce n'est qu'une estimation que nous espérons la plus juste possible. Ce n'est donc pas parce que le coefficient estimé dans le modèle de régression est différent de 0 que le coefficient β de la population est nécessairement

13. La variance σ^2 est un indicateur de la varaibilité dans les données par rapport au modèle optimal.

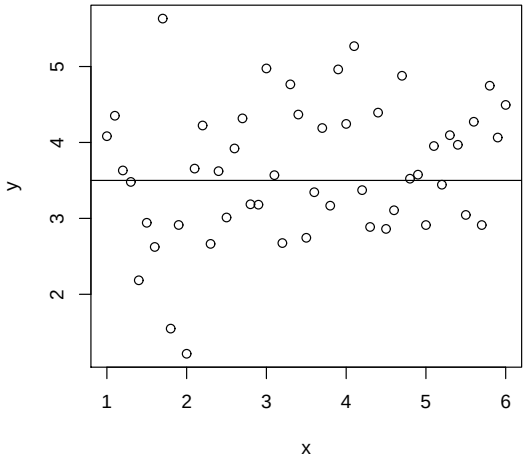

FIGURE 2.3.: Diagramme de dispersion de 51 observations imaginaires, avec une droite de régression ayant pour équation $y = 3.5 + 0 \times x$.

différent de 0. Il est donc souhaitable de savoir si le coefficient β de la population est bien différent de 0. Pour cela, il est d'usage d'effectuer un test t de Student [14].

Nous souhaitons tester l'hypothèse selon laquelle le coefficient β que nous avons calculé est significativement différent de 0. Nous posons l'hypothèse nulle H_0 selon laquelle l'échantillon dans lequel nous avons calculé β est un échantillon appartenant à une population pour laquelle $\beta_{pop} = 0$. Nous allons tenter de rejeter cette hypothèse à l'aide du test t de Student. Dans le cas d'un coefficient de pente, la statistique t se calcule de la façon suivante :

$$t = \frac{\beta - \beta_{pop}}{\sigma(\beta)} \tag{2.4}$$

avec β_{pop} le coefficient de pente dans la population, β le coefficient de pente calculée dans l'échantillon et $\sigma(\beta)$ l'erreur type de β. Si H_0 est vraie, cette statistique suit une loi t de Student avec un degré de liberté égal au nombre d'observations moins deux [15]. Pour les données *cars*, nous avions $\beta = 0.74$. Nous obtenons donc $t = \frac{0.74-0}{0.079} = 9.367$. Avec un degré de liberté égal à 48, cette valeur de t a une *p-value* inférieure à 0.05 ($p < 0.05$). Nous décidons donc de rejeter H_0 et nous concluons que l'échantillon *cars* n'a pas été extrait d'une population pour laquelle $\beta_{pop} = 0$. En d'autres termes, le coefficient β est significativement différent de 0 : les variables `Distance` et `Vitesse` sont bien dépendantes.

14. Pour une présentation de ce test, voir le chapitre 7 de Howell (1998).

15. La distribution t de Student ainsi que la notion de degré de liberté sont présentées dans le chapitre 7 de Howell (1998).

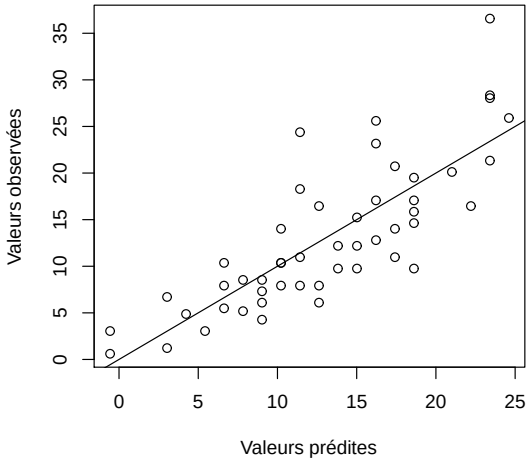

FIGURE 2.4.: Graphique représentant la corrélation entre les valeurs prédites et les valeurs observées pour le modèle linéaire construit sur les données *cars*.

Ainsi, en utilisant le test t de Student, nous pouvons avoir une idée de la significativité du coefficient β et plus précisément, nous pouvons statuer sur le fait qu'il est significativement différent de 0 ou non. C'est le test t qui permet de passer de l'échantillon à la population, en généralisant la relation observée dans l'échantillon.

Qualité de prédiction La méthode d'évaluation de la qualité d'un modèle s'appuie sur la comparaison entre les prédictions faites par le modèle et les données observées. Autrement dit, les valeurs prédites $\hat{y}$ sont comparées aux valeurs observées y, pour chaque valeur de x. L'idée est que plus les valeurs prédites sont corrélées aux valeurs observées, meilleur est le modèle. Pour avoir une idée de la corrélation, nous observons graphiquement les valeurs prédites en fonction des valeurs observées. Pour le modèle linéaire construit sur les données *cars*, nous obtenons le graphique présenté dans la figure 2.4. La droite sur ce graphique représente une corrélation parfaite. Ainsi, plus les points sont groupés le long de cette droite, plus la corrélation est forte et plus le modèle est de bonne qualité.

À cette méthode graphique s'ajoute une méthode numérique. Il s'agit du coefficient de détermination multiple qui est représenté par le symbole R^2. Le principe de cette méthode est d'évaluer la taille des résidus par rapport aux résidus d'un modèle nul, un modèle n'ayant aucun pouvoir prédictif. Le modèle nul ne contient aucune variable prédictrice. La variable à prédire est alors égale à une valeur constante. Plus précisément, pour n'importe quelle valeur de x, ce modèle prédit que y est égal à la moyenne des valeurs de y dans l'échantillon, c'est-à-dire un modèle de la forme : $y = \bar{y}$. Soit A le modèle construit à partir des données et qui est de la forme $y = \alpha + \beta x$, et soit B

le modèle nul, de la forme $y = \bar{y}$. L'idée du coefficient R^2 est de mesurer à quel point le modèle A est meilleur que le modèle B. Pour cela, il faut comparer les résidus, ou erreurs de prédiction, des deux modèles. Rappelons que les erreurs de prédiction d'un modèle se calculent comme la différence des valeurs observées de la variable à prédire et des valeurs prédites : $\epsilon_i = y_i - \hat{y}$. Pour le modèle B, les erreurs s'expriment comme la différence des valeurs observées avec la moyenne $\epsilon_i = y_i - \bar{y}$. La statistique R^2 se calcule en fonction du rapport de la somme des carrés des erreurs du modèle A (SCE_A) sur la somme des carrés des erreurs du modèle B (SCE_B) :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SCE_A}{SCE_B} \tag{2.5}$$

La valeur de R^2 est comprise entre 0 et 1. Plus la valeur est élevée, meilleure est la qualité des prédictions du modèle. Dans le pire des cas, si $SCE_A = SCE_B$, le rapport est égal à 1 et R^2 vaut 0. Plus SCE_A est petit par rapport à SCE_B, plus le rapport va être faible et donc plus R^2 va être élevé. Enfin, si le modèle A représente parfaitement les données, $SCE_A = 0$, ce qui implique que le rapport est égal à 0 et que R^2 vaut 1.

Pour les données *cars*, nous représentons les droites de régression correspondant au modèle A ($Distance = -5.36 + 0.74 \times Vitesse$) et au modèle B ($Distance = \bar{y} = 13.1$) dans le diagramme de dispersion. Dans la figure 2.5, la droite rouge représente le modèle A et la droite verte représente le modèle B. Le calcul de R^2 revient à estimer si la droite de régression rouge résume mieux les données que la droite de régression verte. Le coefficient de détermination multiple est alors : $R^2 = 0.6511$. Cela signifie que le modèle A explique beaucoup mieux la variation de la variable y que le modèle B. Cependant, le modèle A n'est pas complètement satisfaisant car ses erreurs de prédiction sont encore importantes. C'est ce qu'exprime cette valeur de R^2 largement inférieure à 1.

Le cas de régression que nous venons de présenter avec les données *cars* est un cas simplifié. La plupart des cas réels qui sont modélisés impliquent un grand nombre de variables prédictrices. La modélisation se fait alors grâce à la régression multiple, une régression linéaire à plusieurs variables prédictrices.

2.2.1.2. Régression multiple

La régression linéaire multiple est une régression linéaire dans laquelle on modélise la variable à prédire en fonction de n variables prédictrices. Pour cela, le modèle linéaire à une variable est étendu à n variables. Le modèle de régression multiple est donc de la forme :

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_n x_{in} + \epsilon_i \tag{2.6}$$

où

- y_i représente la valeur prédite pour la variable à prédire,
- ϵ_i l'erreur de prédiction,

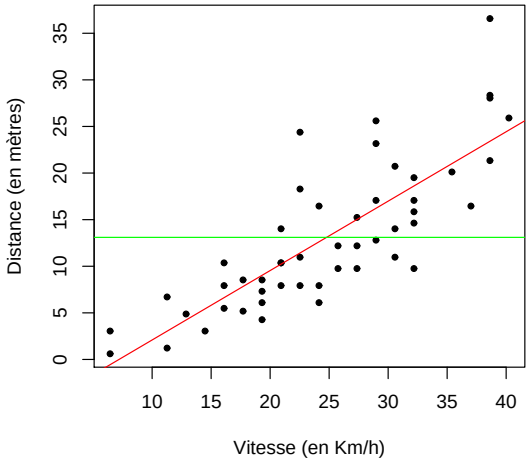

FIGURE 2.5.: Diagramme de dispersion des données *cars*, avec la droite de régression du modèle A en rouge et celle du modèle B en vert.

- x_{ij} la valeur de la j^{eme} variable prédictrice,
- β_j le coefficient de régression partiel associé à la variable x_j.

Comme nous l'avons vu pour le modèle à deux paramètres, pour estimer les paramètres α, β_1, β_2 ...β_n, il faut minimiser la somme des résidus carrés. La méthode des moindres carrés est donc appliquée pour obtenir l'estimation de β_j et de α [16] :

$$\alpha = \bar{y} - \beta_1\bar{x}_1 - \beta_2\bar{x}_2 - \ldots - \beta_n\bar{x}_{jn}$$
$$\beta_j = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$$

Pour illustrer la régression multiple, nous utilisons les données *lexdec*. Ces données rassemblent les temps de décision lexicale de 21 sujets pour 79 noms concrets de l'anglais. L'expérience est la suivante : une suite de lettres est présentée au sujet qui doit décider si ces lettres forment un mot de l'anglais. La donnée expérimentale recueillie est le temps de latence entre le moment où le mot apparaît et le moment où le sujet prend sa décision. Un extrait de cette table de données est présenté dans la table 2.2 [17].

16. Dans le cas de la régression multiple, l'estimation des paramètres revient à trouver les valeurs de α, β_1, β_2 ...β_n, pour lesquelles la fonction dérivée de $SC(\alpha, \beta_1, \beta_2 \ldots \beta_n)$ est nulle. À mesure que le nombre de prédicteurs augmente, la solution à un tel problème devient de plus en plus complexe et relève de méthodes d'optimisation numérique qui ne nous intéressent pas ici.

17. L'utilisation des données *lexdec*, leur modélisation et les représentations graphiques utilisées sont très largement inspirées des cours de Florian Jaeger, disponibles sur : http://www.hlp.rochester.edu/.

	RT	Frequency	Length	NativeLanguage	Subject	Word
1	6.34	4.86	3	English	A1	owl
2	6.31	4.61	4	English	A1	mole
3	6.35	5.00	6	English	A1	cherry
4	6.19	4.73	4	English	A1	pear
5	6.03	7.67	3	English	A1	dog
6	6.18	4.06	10	English	A1	blackberry

TABLE 2.2.: Extrait de la table de données *lexdec*.

Dans un premier temps, nous utilisons les variables `RT`, `Frequency` et `Length` :
- `RT` présente les temps de décision lexicale en millisecondes (échelle logarithmique) ;
- `Frequency` représente la fréquence des lemmes dans la base de données lexicales CELEX (échelle logarithmique) ;
- `Length` représente la longueur des mots en nombre de lettres.

Dans le cas de ces données, nous cherchons à prédire le temps de décision lexicale en fonction de la fréquence du mot et de sa longueur. Il faut donc modéliser la variable à prédire en fonction de deux variables prédictrices. Cela revient à estimer α, β_1 et β_2 dans l'équation :

$$\begin{aligned} RT = \ & \alpha \\ & + \beta_1 Frequency \\ & + \beta_2 Length \\ & + \epsilon \end{aligned}$$

La méthode utilisée est à nouveau celle des moindres carrés et elle permet d'obtenir la solution suivante :

$$\begin{aligned} RT = \ & + 6.51 \\ & - 0.04 \times Frequency \\ & - 0.01 \times Length \\ & \epsilon \sim N(0, 0.235) \end{aligned}$$

Graphiquement, le problème peut être représenté dans un espace à trois dimensions où chaque variable définit un axe. L'équation de régression ne définit alors pas une droite mais un hyperplan de régression qui résume la variable `RT` en fonction de `Frequency` et `Length`. La figure 2.6 présente le diagramme de dispersion des variables `RT`, `Frequency` et `Length` accompagné du plan de régression en orange.

Au-delà de trois variables, la représentation graphique devient compliquée. Néanmoins, la méthode d'estimation des coefficients reste la même. Ainsi, nous ajoutons

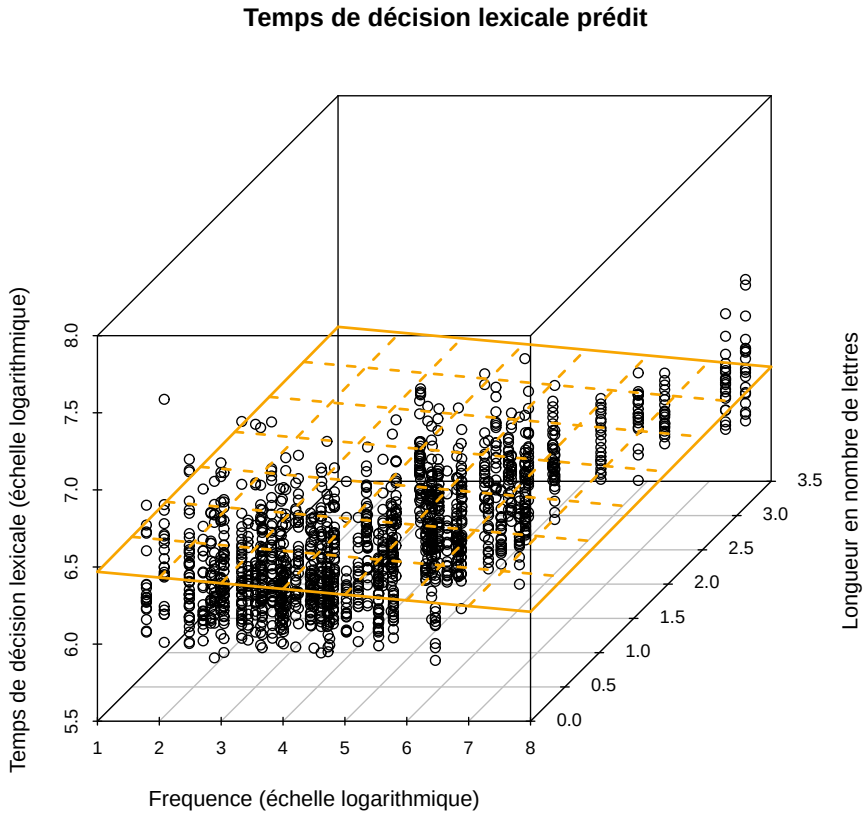

FIGURE 2.6.: Diagramme de dispersion du temps de décision lexicale en fonction de la fréquence et de la longueur, avec la plan de régression en orange (données *lexdec*).

une variable supplémentaire au modèle précédent : `NativeLanguage`. Cette variable indique si le sujet est de langue maternelle anglaise ou non. C'est donc une variable binaire, c'est-à-dire une variable discrète non-ordonnée ayant deux valeurs possibles. La variable `NativeLanguage` prend la valeur 0 si la langue maternelle du sujet est l'anglais et 1 si c'est une autre langue [18]. Le modèle obtenu est présenté dans la figure 2.7.

2.2.1.2.1. Évaluation du modèle Le modèle de régression multiple construit sur l'échantillon doit être évalué. Nous utilisons les mêmes méthodes que celles présentées pour la régression simple.

18. Nous reviendrons plus en détail sur les variables binaires et leur codage dans la partie concernant la régression logistique (partie 2.2.2).

$$
\begin{aligned}
RT = \; & +6.44 \\
& -0.04 \times Frequency \\
& -0.01 \times Length \\
& +0.16 \times NativeLanguage \\
& \epsilon \sim N(0, 0.222)
\end{aligned}
$$

FIGURE 2.7.: Modèle 2

Test d'hypothèse sur les coefficients β_n Nous cherchons à savoir si chaque coefficient de régression est significativement différent de 0, c'est-à-dire si la relation linéaire observée entre chaque variable prédictrice et la variable à prédire peut être généralisée à la population.

Comme nous l'avons vu dans la partie 2.2.1.1.1, c'est le test t de Student qui est appliqué à l'ensemble des coefficients de pente de l'équation du modèle. Dans le cas du Modèle 2, nous obtenons que :

- pour β_1 associé à `Frequency` : $t(1655) = -7.826$ ($p < 0.05$)
- pour β_2 associé à `Length` : $t(1655) = 2.878$ ($p < 0.05$)
- pour β_3 associé à `NativeLanguage` : $t(1655) = 14.155$ ($p < 0.05$)

D'après ces tests, nous concluons que les trois coefficients de régression partiels, β_1, β_2 et β_3, sont significativement différents de 0.

Qualité de prédiction Pour évaluer les capacités de prédiction de ce modèle de régression, nous représentons, dans la figure 2.8, la corrélation entre les temps de décision lexicale prédits et les temps de décision lexicale observés. Ce graphique indique que la corrélation n'est pas très bonne car les points sont très dispersés autour de la droite de corrélation parfaite. Ce résultat graphique est confirmé par la valeur du coefficient de détermination multiple (cf. partie 2.2.1.1.1), $R^2 = 0.15$. Cette valeur indique que le modèle 2 n'est pas vraiment meilleur qu'un modèle qui prédirait systématiquement que le temps de décision lexicale est égal à la moyenne des temps observés dans les données *lexdec*.

Surentraînement Dans certains cas, il arrive que le modèle soit surentraîné sur les données de l'échantillon. Cela signifie que le modèle calculé sur l'échantillon a de très bonnes performances sur les données de l'échantillon, mais il est beaucoup moins bon sur des données inconnues. Le modèle est, en quelque sorte, collé aux données et ne présente donc pas de bonne capacité de généralisation. Il existe plusieurs méthodes pour vérifier le surentraînement (*overfitting*). Nous présentons ici celle que nous utiliserons : la validation croisée (*cross-validation*). L'idée est de différencier

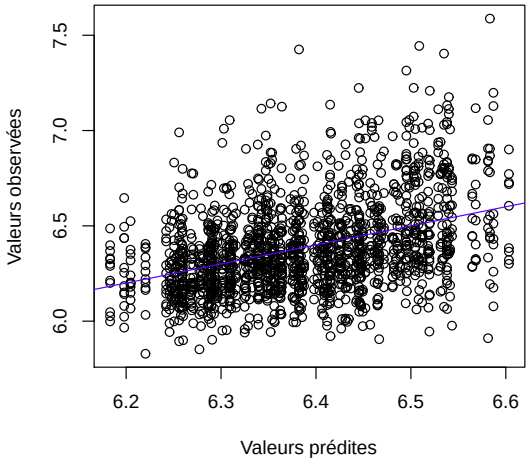

FIGURE 2.8.: Graphique de la corrélation entre les temps de décision lexicale prédits par le Modèle 2 et les temps observés dans les données *lexdec*. La droite indique une corrélation parfaite.

les données d'entraînement et les données de test du modèle. Pour cela, le corpus est divisé en une partie entraînement et une partie test. La qualité du modèle est alors évaluée sur des données inconnues, ce qui évacue tout problème de surentraînement. Afin de tester le modèle sur l'ensemble des données, nous utilisons une validation croisée à k passes. Dans ce cas, le corpus est divisé k fois en échantillons d'entraînement et échantillons de test, et le modèle est construit puis évalué sur les k couples entraînement/test. Ainsi, l'évaluation du modèle est opérée k fois sur des données inconnues. Si le coefficient de détermination multiple R^2 est calculé sur chaque échantillon de test, c'est la moyenne μ des R^2 obtenus ainsi que l'écart type σ qui sont reportés. Pour illustrer cela, nous utilisons le Modèle 2 construit sur les données *lexdec*. Ce modèle a un coefficient de détermination multiple : $R^2 = 0.15$. Ce coefficient montre que le Modèle 2 n'explique pas vraiment mieux la variation des données qu'un modèle prédisant systématiquement la moyenne des temps de décision lexicale. Nous faisons une validation croisée à 100 passes et nous obtenons : $\mu = 0.15$. Cette moyenne indique qu'il n'y a pas de surentraînement au niveau du Modèle 2.

La régression multiple permet de modéliser une variable en fonction de n variables prédictrices. Le nombre de variables est à déterminer et il existe une méthode permettant de sélectionner les variables prédictrices selon qu'elles participent significativement ou non à la modélisation de la variable à prédire. L'objectif est alors de trouver le modèle le plus compact, c'est-à-dire le meilleur modèle contenant le moins de variables prédictrices.

2.2.1.2.2. Comparaison de modèles : trouver le modèle le plus compact Nous pouvons nous interroger sur la capacité de chaque variable à améliorer la prédiction de la variable dépendante. Pour aborder cette question, nous nous intéressons à la sélection de variables, ce qui est traditionnellement fait à l'aide de la statistique F. Cette statistique correspond à un rapport de variance dont la distribution suit une loi de Fisher Snedecor [19]. Pour sélectionner les variables qui contribuent significativement au modèle, il faut comparer différents modèles qui sont imbriqués. Soient le Modèle G et le Modèle P, les modèles à comparer, le Modèle G doit contenir l'ensemble des variables prédictrices du Modèle P plus une au moins. La comparaison se fait généralement entre un Modèle G contenant n variables prédictrices, et un Modèle P contenant $n-1$ variables prédictrices. L'objectif de la comparaison de ces deux modèles est de statuer sur le pouvoir explicatif de la variable présente dans le Modèle G et pas dans le Modèle P. Plus largement, la méthode de sélection des variables permet de compacter le modèle de régression, c'est-à-dire de trouver le modèle ayant le meilleur pouvoir explicatif et contenant le moins de variables prédictrices.

La comparaison des Modèles G et P se fait également à partir de la somme des carrés des erreurs de prédiction des deux modèles : SCE_G et SCE_P. Un autre élément est à prendre en compte : celui du nombre de variables prédictrices, puisque l'objectif est de trouver le modèle le plus compact. Soient n_G et n_P les nombres de variables prédictrices respectivement des Modèles G et P, et soit m le nombre d'observations, la statistique F s'exprime de la façon suivante :

$$F = \frac{\frac{SCE_P - SCE_G}{n_G - n_P}}{\frac{SCE_G}{m - n_G}} \tag{2.7}$$

Cette statistique met en rapport deux variances : d'abord, la variance de la différence des résidus entre les deux modèles ($\frac{SCE_P - SCE_G}{n_G - n_P}$), autrement dit la part de la variance des données qui est expliquée par le Modèle G et pas par le Modèle P ; puis la variance des résidus du Modèle G, le modèle le plus complexe ($\frac{SCE_G}{m - n_G}$), autrement dit la part de la variance des données qui ne peut pas être expliquée par le Modèle G. Si le Modèle G et le Modèle P sont quasi équivalents, la variance expliquée par le Modèle G sans être expliquée par le modèle P est très faible. La statistique F sera alors faible. Plus la variance non expliquée par le Modèle P mais expliquée par le Modèle G est grande, plus la valeur du rapport va avoir tendance à augmenter.

Etant donné que la statistique F est distribuée selon une loi de Fisher Snedecor, un test d'hypothèse reposant sur cette distribution est appliqué pour décider si la statistique F a une valeur assez importante pour dire que le Modèle G est meilleur que le Modèle P. L'hypothèse nulle H_0 se définit alors de la façon suivante : le Modèle G n'est pas meilleur que le Modèle P pour expliquer les données. Sous cette hypothèse, la statistique F a une distribution de Fisher avec les degrés de liberté $n_G - n_P$ et $m - n_G$. L'hypothèse nulle est rejetée si la probabilité que la statistique F calculée

19. Pour une présentation de la statistique F et de la loi de Fisher, se reporter au chapitre 4 de Judd *et al.* (2010).

appartienne à la distribution de F, étant donné H_0, est inférieure à 0.05. Si H_0 est rejetée, alors nous estimons que le Modèle G est meilleur que le Modèle P pour expliquer les données. En d'autres termes, en cas de rejet de H_0, nous concluons que la variable contenue dans le Modèle G et absente du Modèle P a un pouvoir explicatif.

A titre d'illustration, nous présentons la comparaison des Modèles 1 et 2 en utilisant la statistique F pour évaluer le pouvoir explicatif de la variable `NativeLanguage`. En effet, le Modèle 1 est imbriqué dans le Modèle 2 et se différencie de ce dernier par une seule variable : `NativeLanguage`. Pour le calcul de la statistique F, le Modèle 1 correspond au Modèle P et le Modèle 2 au Modèle G. La statistique F est alors égale à 200.5. Or, cette valeur de F avec des degrés de liberté de 1 et 1656 a une *p-value* proche de 0 ($p< 2.16 \times 10^{-16}$). La statistique F permet donc de rejeter l'hypothèse nulle selon laquelle le Modèle 1 est aussi bon que le Modèle 2 pour expliquer les données. En conclusion, la variable `NativeLanguage` permet d'expliquer une part significative de la variance des données.

2.2.1.2.3. Interprétation des coefficients Une fois le modèle de régression multiple obtenu, il serait intéressant de pouvoir interpréter les coefficients pour donner du sens au modèle. Nous cherchons à déterminer l'influence de chaque variable dans le modèle. Par exemple, dans le Modèle 2, nous nous demandons si les variables `Frequency`, `Length` et `NativeLanguage` ont tendance à faire augmenter ou diminuer le temps de décision lexicale. Nous souhaitons également évaluer l'importance relative de l'effet de chaque variable sur le phénomène étudié : est-ce que la variable `Frequency` a un effet plus important que la variable `Length` ?

De façon schématique, un coefficient positif indique que la variable prédictrice favorise une augmentation de la valeur de la variable à prédire, et inversement, un coefficient négatif signale que la variable prédictrice a pour influence la diminution de la valeur de la variable à prédire. Notons que si la variable prédictrice a pour valeur des réels positifs et négatifs, l'influence de la variable dépend de son propre signe. Dans le Modèle 2, nous observons que les variables `Frequency` et `Length` ont un coefficient négatif, ce qui indique qu'elles favorisent la diminution de la variable à prédire. En d'autres termes, plus la fréquence et la longueur du mot augmentent, plus le temps de décision lexicale a tendance à diminuer. La variable `NativeLanguage`, pour sa part, a un coefficient positif, ce qui signifie que, quand la variable prédictrice vaut 1, elle fait augmenter la valeur de la variable à prédire. Autrement dit, lorsque le sujet est de langue maternelle non anglaise, le temps de décision lexicale a tendance à augmenter. Cependant, deux problèmes majeurs se posent pour une interprétation plus approfondie des coefficients : la corrélation et l'échelle des variables.

La corrélation des variables Pour présenter le problème qui nous intéresse ici, citons un passage de l'article de Baayen *et al.* (2006) qui expose clairement les enjeux de la corrélation dans la régression : « *Lorsque les prédicteurs d'une régression multiple ne sont pas corrélés, chaque prédicteur rend compte d'une partie unique de la variance. En d'autres termes, lorsqu'il n'y a pas de colinéarité, la valeur explicative de chaque*

prédicteur pris individuellement peut être estimée correctement. Cela n'est pas possible quand les prédicteurs sont colinéaires. Ensemble, les variables colinéaires peuvent expliquer presque toute la variance, mais il n'est pas possible de savoir quelle part de variance est expliquée par quelle variable. » [20]

Un premier cas de corrélation peut s'observer entre deux variables prédictrices. Dans le cas de variables continues, nous utilisons une mesure de corrélation, telle que le coefficient de corrélation de Pearson [21] qui permet de détecter une corrélation linéaire. Par exemple, pour les variables `Frequency` et `Length`, dans les données *lexdec*, $r = -0.429$. Le signe de r indique que la corrélation est négative : la baisse de la longueur est corrélée avec l'augmentation de la fréquence. La valeur du coefficient indique qu'il existe une corrélation mais qu'elle est relativement faible. Dans le cas de variables binaires, on peut évaluer l'indépendance de deux variables à l'aide d'un test de χ^2 [22].

Le problème de la colinéarité peut être plus complexe. On parle alors de multicolinéarité. Dans ce cas, une variable est corrélée à plusieurs autres variables. Plus précisément, une variable prédictrice x_i est corrélée à la combinaison linéaire d'autres variables prédictrices $\beta_j x_j + \beta_k x_k + \ldots + \beta_n x_n$. Nous présentons ici deux méthodes qui permettent d'évaluer la multicolinéarité des variables.

La première mesure permettant d'évaluer la multicolinéarité est l'indice de conditionnement (*condition number*, κ). Cette mesure unique donne une image générale de la multicolinéarité des variables prédictrices d'un modèle [23]. Une valeur κ inférieure à 6 indique qu'il n'y a pas de multicolinéarité dans le modèle. Autour de 15, le κ signale un niveau de colinéarité moyen. Enfin, un κ supérieur à 30 révèle une colinéarité problématique. Pour le calcul de l'indice de conditionnement, nous suivons la méthode proposée par Belsley *et al.* (1980) et implémentée en R par Baayen (2008) dans la fonction `collin.fnc`. Dans le cas du Modèle 2, nous obtenons une valeur $\kappa = 14.954$ [24]. Cette valeur indique que les variables prédictrices présentent

20. « *When the predictors in multiple regression are uncorrelated, each predictor accounts for a unique portion of the variance. In other words, when there is no collinearity, the explanatory value of each individual predictor can be properly assessed. This is not possible when the predictors are collinear. Together, collinear variables may explain nearly all the variance, but it is never clear what part of the variance is explained by which variable* » (Baayen *et al.*, 2006, 295)

21. Le coefficient de corrélation de Pearson prend ses valeurs dans l'intervalle [-1,1]. Un coefficient r de valeur 0 indique que les variables sont linéairement indépendantes. Plus la valeur du coefficient s'approche de 1 ou de -1, plus la relation de corrélation linéaire entre les deux variables est forte. Le signe du coefficient indique le sens de la corrélation.

22. Le test du *chi*^{2} est présenté dans le chapitre 6 de Howell (1998).

23. Pour plus de détails sur le calcul de l'indice de conditionnement, voir Belsley *et al.* (1980).

24. Il faut remarquer que le calcul de l'indice de conditionnement dans le cas de variables nominales, telles que `NativeLanguage`, ne va pas de soi. Comme l'expliquent Wissman *et al.* (2007), « *In linear regression analysis, the dummy variables also play an important role as a possible source of multicollinearity. The choice of reference category for a categorical variable may affect the degree of multicollinearity in the data* ». Dans le cadre de cette thèse, nous utilisons des variables nominales binaires et la catégorie de référence correspond à la valeur de la variable à laquelle on attribue 0. Pour la variable `NativeLanguage`, la catégorie de référence est `English`. Notons que la valeur de l'indice de conditionnement est légèrement supérieure lorsque nous utilisons `Other` comme catégo-

une multicolinéarité moyenne.

Pour établir plus précisément quelles sont les variables problématiques, une deuxième mesure est utilisée : le facteur d'inflation de la variance (*Variation Inflation Factor*, VIF). Le VIF permet d'évaluer à quel point le coefficient associé à une variable a une variance élevée en raison du degré de corrélation de la variable avec les autres variables prédictrices. Le calcul du VIF se fait pour chaque variable prédictrice du modèle de régression $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \epsilon$. L'idée est de calculer une régression linéaire pour chaque variable prédictrice x_i en fonction de toutes les autres variables prédictrices, en excluant la variable à prédire y. Cela permet d'évaluer à quel degré la variable x_i est prédictible à partir des autres variables du modèle. Pour cette nouvelle régression, nous pouvons calculer un coefficient de détermination multiple $R^2_{x_i}$. Plus ce coefficient est élevé, plus le facteur x_i est redondant avec les autres variables prédictrices. La tolérance, calculée comme $1 - R^2$, représente, au contraire, la part de variance de la variable x_i qui n'est pas expliquée par les autres variables prédictrices et reflète donc sa singularité. Plus la tolérance est élevée, moins la multicolinéarité est importante. Enfin, un rapport basé sur la tolérance est défini afin d'estimer le degré d'augmentation de la variance du coefficient de la variable x_i dans le modèle de régression $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \epsilon$:

$$VIF(\beta_i) = \frac{1}{1 - R^2_{x_i}} \tag{2.8}$$

Ainsi, pour chaque variable x_i du modèle de régression, nous évaluons le degré de variance du coefficient β_i dû à la corrélation de x_i avec les autres variables prédictrices du modèle. Si la valeur de VIF pour une variable est égale à 1, la variable x_i ne présente aucune corrélation avec les autres variables. Plus le VIF est élevé, plus le risque de colinéarité est important. Il est généralement admis qu'une valeur VIF supérieure à 5 signale une forte multicolinéarité.

Dans le cas du Modèle 2, nous calculons les facteurs d'inflation de la variance pour les coefficients des variables `Frequency`, `Length` et `NativeLanguage` : $VIF(\beta_{\texttt{Freq}}) = 1.23$, $VIF(\beta_{\texttt{Len}}) = 1.23$ et $VIF(\beta_{\texttt{NatLan}}) = 1$. D'après ces facteurs d'inflation, nous observons que la variable `NativeLanguage` ne peut pas être expliquée à partir des deux autres variables ($VIF(\beta_{\texttt{NatLan}}) = 1$), ce qui est tout à fait attendu dans la mesure où la fréquence et la longueur d'un mot sont des dimensions orthogonales à la langue maternelle du sujet d'une expérience. Cette valeur de VIF signifie que la variable ne présente pas de problème de multicolinéarité. Le VIF des deux autres variables est identique puisque la variable `NativeLanguage` n'intervient pas. La valeur de ce VIF est relativement faible, ce qui indique que le Modèle 2 n'est pas sérieusement affecté par la multicolinéarité.

Il existe des méthodes permettant de réduire la multicolinéarité dans un mo-

rie de référence : $\kappa = 15.294$. Le changement de la catégorie de référence d'une variable binaire peut donc être envisagé comme une méthode de réduction de la multicolinéarité pour les variables binaires.

dèle. La première consiste simplement à supprimer une ou plusieurs variables qui apparaissent comme trop redondantes par rapport à d'autres. Cependant, il faut être prudent avec cette méthode, car supprimer une variable peut impliquer la suppression d'une partie d'information pertinente. Nous reprenons ici l'exemple de Belsley *et al.* (1980) qui présentent le cas d'un test évaluant les capacités d'étudiants en mathématiques et en physique. Les résultats en mathématiques et en physique sont généralement très corrélés. Cependant, si l'on sélectionne les étudiants en se fondant uniquement sur le résultat en physique, on éliminera des étudiants qui sont excellents en mathématiques mais qui ne portent aucun intérêt à la physique. En d'autres termes, une corrélation importante entre deux variables x et y n'indique pas forcément que l'on peut prédire avec précision, pour l'ensemble des valeurs de x, une valeur y correcte. La suppression de variables est donc, en général, réservée aux variables qui sont mathématiquement liées et donc pour lesquelles les valeurs de l'une peuvent être obtenues avec précision à partir des valeurs de l'autre. Une autre méthode possible est de combiner mathématiquement deux variables corrélées pour n'obtenir qu'une seule variable. Par exemple, on peut faire le rapport de deux fréquences, comme le proposent Baayen *et al.* (2006), pour la fréquence du mot à l'écrit et la fréquence du mot à l'oral dans leur modèle. Enfin, il existe d'autres méthodes plus complexes (analyse en composantes principales, régression en composantes principales…) que nous n'aborderons pas dans cette thèse.

Nous avons décrit les mesures de corrélation et de mutlicolinéarité que nous utiliserons. L'évaluation et le diagnostic des problèmes de mutlicolinéarité dans les modèles de régression multiple sont capitaux. La réduction maximale de ce problème permet de faciliter l'interprétation de ces coefficients.

L'échelle des variables Lorsque le modèle construit a une multicolinéarité très faible et que nous souhaitons interpréter l'importance relative des coefficients, il faut faire attention à l'échelle des variables. En effet, l'importance de chaque coefficient est relative à l'échelle de mesure de la variable. Si une variable x_a prend ses valeurs dans l'intervalle [0,100], une variable x_b dans l'intervalle [0,1] et que ces deux variables se voient attribuer le même coefficient dans un modèle de régression linéaire, cela ne signifie pas que ces deux variables ont le même effet sur la variable à prédire. En effet, dans un cas le coefficient peut être multiplié par 100 alors que dans l'autre cas, il l'est au maximum par 1. Pour avoir une meilleure image de l'apport relatif de chaque variable, celles-ci peuvent être mises à la même échelle. Dans le cas des variables continues distribuées normalement, la méthode traditionnelle consiste à standardiser les valeurs autour d'une moyenne 0 et d'un écart type de 1. Pour cela, la distribution de chaque variable est transformée en une distribution z. Soient x la variable à transformer, $\bar{x}$ sa moyenne et σ son écart type, le score z se définit de la façon suivante :

$$z = \frac{x - \bar{x}}{\sigma} \tag{2.9}$$

Les coefficients de régression associés à ces variables standardisées pourront alors être comparés sans problème d'échelle. Pour le Modèle 2, on utilise les scores z de `RT`, `Frequency` et `Length`. Le modèle de régression obtenu est présenté dans la figure 2.9. Les valeurs des coefficients $\beta_{\texttt{Freq}}$ et $\beta_{\texttt{Len}}$ sont supérieures aux coefficients du

$$\begin{aligned} RT = \quad & -0.27 \\ & -0.19 \times Frequency \\ & -0.07 \times Length \\ & +0.65 \times NativeLanguage \\ & \epsilon \sim N(0, 0.222) \end{aligned}$$

FIGURE 2.9.: Modèle 2 bis

Modèle 2 (respectivement -0.04 et -0.01). Nous pouvons estimer que, dans le cadre de ce modèle, c'est-à-dire lorsque seules les trois variables du modèle sont prises en considération, l'augmentation d'un point du score z de la fréquence fait diminuer le score z du temps de décision lexicale de 0.19, tandis que l'augmentation d'un point du score z de la longueur fait augmenter de 0.07 le score z du temps de décision lexicale. Dans le Modèle 2, la variable `Frequency` a donc un effet plus important que `Length`, toutes choses égales par ailleurs. Comme nous le voyons dans cet exemple, la standardisation des variables permet de comparer les coefficients, mais ajoute une difficulté dans l'interprétation des variables : à quoi correspond le score z de la longueur ou du temps de décision lexicale ? De plus, il reste un problème pour la comparaison des coefficients : les variables binaires comme `NativeLanguage`. Ces dernières ne peuvent pas être standardisées et ne peuvent donc pas partager la même échelle que les variables continues. La situation peut se résumer ainsi : dans le cadre d'un modèle de régression satisfaisant avec une multicolinéarité faible,

- il est facile de comparer les coefficients de variables binaires dans un modèle car elles partagent les mêmes valeurs (1 et 0) ;
- il est possible de comparer les coefficients de variables continues en les standardisant ;
- mais il est très compliqué de comparer les coefficients d'une variable binaire et d'une variable continue.

Nous avons présenté le modèle de régression linéaire qui permet de décrire le comportement d'une variable en fonction de n variables prédictrices. D'un point de vue méthodologique, le modèle doit être construit en sélectionnant les variables ayant un véritable pouvoir explicatif. Il faut ensuite vérifier la qualité de la régression et les problèmes de multicolinéarité. Un modèle ayant une bonne qualité de prédiction, sans surentraînement, est un bon outil de généralisation. De plus, si les variables

prédictrices ne présentent pas de multicolinéarité, on peut interpréter les coefficients du modèle afin de mieux comprendre le phénomène modélisé.

2.2.1.3. Regression linéaire à effets mixtes

L'amélioration d'un modèle de régression se fait en minimisant la variance non expliquée des données qui n'est pas due au hasard. Dans le cas de données expérimentales, telles que celles de *lexdec*, l'une des sources de variation est le sujet qui a passé l'expérience. En effet, le temps de décision lexicale mesuré dépend largement de la personne qui passe l'expérience : il existe des personnes plus rapides que d'autres, certaines peuvent être plus fatiguées, moins concentrées que d'autres… Si nous observons le temps de décision lexicale moyen de chacun des 21 sujets ayant participé à l'expérience *lexdec*, nous constatons que ces temps moyens sont très variables d'un individu à l'autre. Dans le graphique de droite de la figure 2.10, la moyenne du temps de décision lexicale de chaque individu est représentée par l'index qui désigne le sujet. La droite grisée indique la moyenne générale de temps de décision lexicale dans cette expérience. Les sujets se situant au-dessus de cette droite sont plus lents que la moyenne et ceux qui sont au-dessous sont plus rapides. De plus, si l'on visualise la variable de temps de décision lexicale en fonction de la fréquence des mots, on observe des groupes de données. Dans le graphique de gauche de la figure 2.10, nous avons fait ressortir les temps de trois sujets. Nous voyons que les données relatives à chaque individu forment des groupes dans le diagramme de dispersion. Cela signifie que les données ne sont pas organisées au hasard, elles sont en grande partie structurées, notamment autour de chaque sujet. Ainsi, en plus du phénomène général qui veut que le temps de décision lexicale tende à diminuer pour les mots très fréquents, chaque individu a un comportement singulier dont il faut rendre compte si l'on veut diminuer la variation non exliquée des données. Un première solution pourrait être de faire une droite de régression pour chaque individu. C'est cette idée qui est représentée par la figure 2.11. La droite rouge représente la droite de régression pour le sujet A1, la droite bleue celle de l'individu Z et la verte celle du sujet T2. Les droites de régression mettent en lumière le fait que les données de chaque sujet sont groupées autour de différentes droites de régression. Les trois droites sont orientées à peu près de la même façon, laissant penser qu'il existe bien une corrélation entre temps de lecture et fréquence, au-delà des différences individuelles. Plus précisément, les trois droites se distinguent par leur intercept ($\alpha_{A1} = 6.53$, $\alpha_Z = 7.04$, $\alpha_{T2} = 7.15$), et par leur pente ($\beta_{A1} = -0.05$, $\beta_Z = -0.09$, $\beta_{T2} = -0.07$). Schématiquement, nous pouvons interpréter les différents intercepts comme des différences de réactions individuelles face à la tâche de l'expérience. Les différences de pente mettent en avant le fait que chaque individu ne réagit pas tout à fait de la même façon à la variable fréquence du mot. Le problème soulevé par une telle méthode est la perte de généralisation sur le phénomène. En effet, nous obtenons plusieurs modèles de régression linéaire (autant que de sujets de l'expérience) qui modélisent le comportement de chaque individu, mais qui ne rendent pas compte de façon générale de la question qui nous intéresse : quels éléments exercent une influence sur le temps de décision lexicale chez les adultes ?

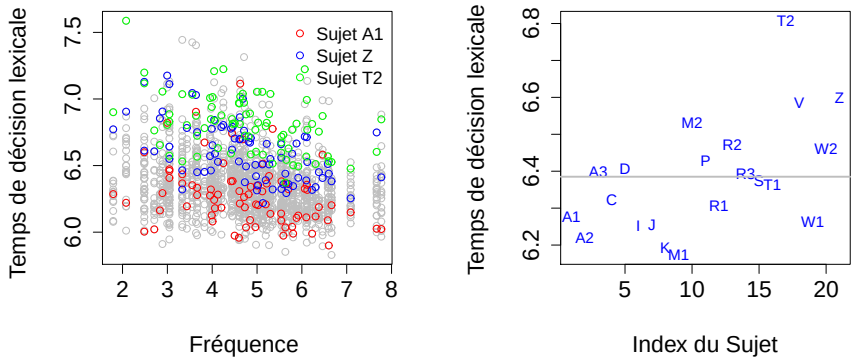

FIGURE 2.10.: À gauche : diagramme de dispersion du temps de lecture en fonction de la fréquence des lemmes *lexdec* avec des points de couleur représentant trois sujets différents. À droite : représentation graphique des temps de décision lexicale moyen par sujet avec la droite grise indiquant la moyenne générale des temps de décision lexicale pour les données *lexdec*

Pour répondre à cette question tout en tenant compte de la structure des données, une méthode possible est la régression linéaire à effets mixtes. Ce type de régression est dit "à effets mixtes" car il contient deux types d'effets : les effets fixes et les effets aléatoires. Les effets fixes correspondent aux effets des variables prédictrices, telles que nous les avons vues dans les modèles linéaires (Modèle 1 et Modèle 2). Les effets aléatoires renvoient aux effets des variables qui forment des groupes dans les données, telles que la variable `Subject` dans les données *lexdec*. Les modèles de régression à effets mixtes se composent donc, premièrement, de variables prédictrices auxquelles sont associés des coefficients comme dans le modèle linéaire, et deuxièmement, de coefficients affectés à chaque valeur possible des effets aléatoires. Par exemple, pour les données *lexdec*, nous modélisons le temps de décision lexicale (`RT`) en fonction de la variable `Frequency`, tout en tenant compte de la structuration des données autour de la variable `Subject`. L'objectif est de rendre compte du fait que `Frequency` influe de façon générale sur `RT`, que chaque individu a en moyenne un `RT` qui varie et que chaque individu a une sensibilité différente à la variable `Frequency`. `Frequency` est traité comme un effet fixe ayant un coefficient de pente. `Subject` est traité comme un effet aléatoire et chaque valeur de la variable `Subject` se voit attribuer un intercept aléatoire et un coefficient de pente associé à la variable `Frequency`. L'intercept aléatoire spécifique à chaque sujet rend compte d'une différence générale entre les sujets, certains étant plus rapides que d'autres. La pente aléatoire spécifique à chaque sujet rend compte des différences de réaction par rapport à la fréquence, certains sujets étant plus sensibles que d'autres à la variable `Frequency`. Le modèle ainsi décrit se formalise comme cela est présenté dans la figure 2.12.

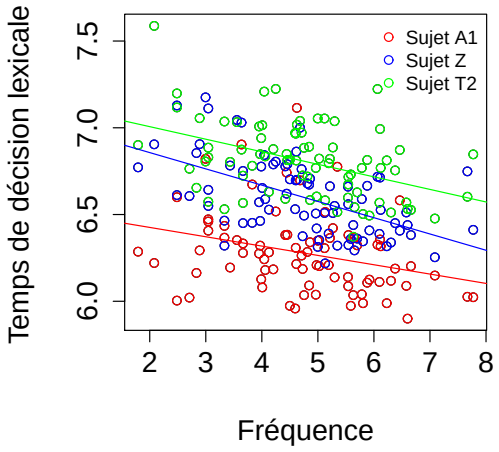

FIGURE 2.11.: Diagramme de dispersion du temps de lecture en fonction de la fréquence des lemmes pour trois sujets (*lexdec*). Les trois droites de couleur correspondent à la régression pour les sujets A1, Z et T2.

$$\begin{aligned} RT_{\mathbf{subject=i}} = \; & \alpha + \alpha_{\mathbf{subject=i}} \\ & + (\beta + \beta_{\mathbf{subject=i}}) \times Frequency \\ & + \epsilon \end{aligned}$$

où
- ϵ représente l'erreur de prédiction,
- α l'intercept général du modèle,
- β le coefficient de pente de la variable `Frequency`,
- $\alpha_{subject=i}$ l'intercept aléatoire associé à la valeur i de la variable `Subject`,
- $\beta_{subject=i}$ le coefficient de pente de `Frequency` associé à la valeur i de la variable `Subject`.

FIGURE 2.12.: Modèle 3

Chaque valeur de la variable `Subject` se voit attribuer un intercept et une pente propres qui s'ajoutent linéairement aux intercept et pente du modèle général et qui permettent en quelque sorte de définir une droite de régression pour chaque individu. Pour les sujets A1, T2 et Z qui apparaissent dans la figure 2.11, nous obtenons les trois équations suivantes :

- $RT_{A1} = (6.5888 - \mathbf{0.1189}) + (-0.0429 + \mathbf{0.0033}) \times Frequency + \epsilon, \epsilon \sim N(0, 0.0321)$
- $RT_{T2} = (6.5888 + \mathbf{0.5961}) + (-0.0429 - \mathbf{0.0384}) \times Frequency + \epsilon, \epsilon \sim N(0, 0.0321)$

– $RT_Z = (6.5888 + \mathbf{0.3655}) + (-0.0429 - \mathbf{0.0320}) \times Frequency + \epsilon, \epsilon \sim N(0, 0.0321)$

Les intercepts aléatoires indiquent que A1 est légèrement plus rapide que la moyenne tandis que T2 et Z sont un peu plus lents. Les pentes aléatoires signalent que l'effet de la fréquence est un peu atténué chez le sujet A1 alors qu'il est légèrement renforcé chez les sujets T2 et Z. La distribution des valeurs des intercepts et des pentes associées aux 21 sujets de l'expérience, peut être observée sur la figure 2.13. Ces intercepts et pentes

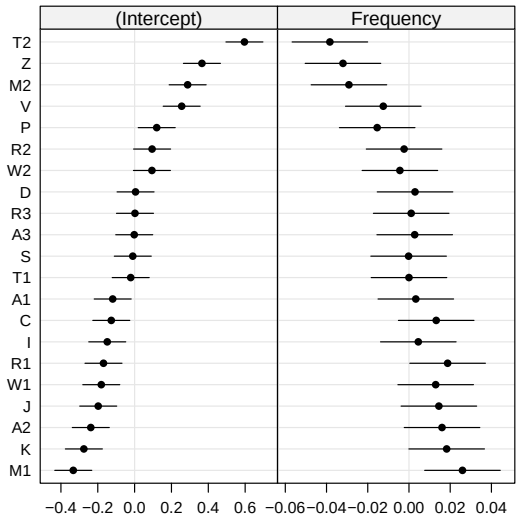

FIGURE 2.13.: Distribution des effets aléatoires associés à la variable `Subject` (*lexdec*). La barre horizontale autour de chaque point représente l'intervalle de confiance à 95%

« *sont appelés effets aléatoires car ils sont associés avec des unités expérimentales particulières qui sont sélectionnées au hasard dans la population concernée* » [25]. En effet, les sujets ne représentent qu'un échantillon pris au hasard dans l'ensemble des sujets possibles. De plus, « *ce sont des effets car ils représentent une déviation par rapport à une moyenne générale* » [26]. Une représentation formelle possible des modèles à effets mixtes, avec intercept et pentes aléatoires, est la suivante :

$$y = X\beta + Zb + \epsilon \tag{2.10}$$

où

– y correspond à la variable à prédire ;
– X renvoie à la matrice représentant l'ensemble des variables prédictrices ;
– β à la matrice contenant l'ensemble des effets fixes du modèle (intercept et coefficients de pente partiels) ;

25. « *…are called random effects because they are associated with the particular experimental units that are selected at random from the population* » (Pinheiro & Bates, 2000, p. 8)

26. « *They are effects because they represent a deviation from an overall mean.* » (Pinheiro & Bates, 2000, p. 8)

– Z renvoie à la matrice représentant l'ensemble des variables à effets aléatoires ;
– b à la matrice contenant l'ensemble des valeurs des effets aléatoires.

Tout comme les résidus, les effets aléatoires sont définis comme des variables aléatoires distribuées normalement ayant une moyenne égale à 0 et un écart type à définir. Un effet aléatoire x se note donc : $x \sim N(0, \sigma_x)$. Comme l'expliquent Baayen *et al.* (2008), « *quand un modèle à effets mixtes est ajusté à un ensemble de données, son ensemble de paramètres estimés comprend, d'un côté, les coefficients pour les effets fixes, et, de l'autre côté, les écarts-types et les corrélations pour les effets aléatoires. Les valeurs particulières des ajustements effectués pour les intercepts et les pentes sont calculées une fois que les paramètres aléatoires ont été estimés. Formellement, ces ajustements, appelés* Best Linear Unbiased Predictors *(or* FBLUPS*), ne sont pas des paramètres du modèle.* » [27]

L'estimation des paramètres β, b et ϵ d'un modèle se fait généralement à l'aide du maximum de vraisemblance (*Maximum Likelihood*, **ML**). Cette méthode suit un principe analogue à la méthode des moindres carrés que nous avons vue pour la régression linéaire. L'idée est de trouver les valeurs des paramètres qui permettent de maximiser la fonction de vraisemblance de données (Gelman & Hill, 2007). L'estimation des paramètres s'effectue de façon itérative à l'aide d'algorithmes d'optimisation numérique qui ne nous intéressent pas dans le cadre de notre travail. Nous utilisons les résultats des algorithmes implémentés dans l'extension `lme4` de R (Bates & Sarkar, 2007).

Une fois le modèle obtenu, son évaluation se fait comme pour la régression linéaire multiple en statuant sur la significativité des coefficients et en mesurant la qualité des prédictions faites par le modèle. Nous n'entrons pas dans les détails en ce qui concerne le modèle linéaire à effets mixtes. En revanche, nous donnons quelques éléments concernant la sélection des variables pour obtenir le modèle le plus compact.

2.2.1.3.1. Comparaison de modèles : trouver le modèle le plus compact La sélection des variables dans un modèle linéaire à effets mixtes suit exactement les mêmes principes méthodologiques que pour le modèle linéaire vu précédemment. Cependant, étant donné que la construction du modèle s'appuie sur des méthodes de calcul différentes de celles utilisées pour le cas sans effet aléatoire, la méthode de sélection ne peut pas se faire sur la base de la somme des carrés des erreurs (SCE).

La méthode utilisée pour comparer deux modèles enchâssés est le test du rapport de vraisemblance (*Likelihood Ratio Test*). L'idée est de comparer la vraisemblance des deux modèles telle qu'elle a été estimée à l'aide de la méthode ML. Si les deux vraisemblances sont très proches, alors il est estimé que le modèle le plus simple,

27. « *When a mixed-effects model is fitted to a data set, its set of estimated parameters includes the coefficients for the fixed effects on the one hand, and the standards deviations and correlations for the random effects on the other hand. The individual values of the adjustments made to intercepts and slopes are calculated once the random-effects parameters have been estimated. Formally, these adjustments, referenced as Best Linear Unbiased Predictors (*BLUPS*), are not parameters of the model.* »

Modèle P, est préférable ; et si la vraisemblance du modèle plus complexe, Modèle G, est significativement plus importante que celle du Modèle P, alors on conclut que la complexité du Modèle G est justifiée et on conserve le Modèle G. Soient D la statistique du test de rapport de vraisemblance, L_G et L_P les vraisemblances respectives du Modèle G et du Modèle P, on a :

$$D = 2log(\frac{L_G}{L_p}) = 2(log(L_G) - log(L_P)) \tag{2.11}$$

Sous l'hypothèse nulle selon laquelle le modèle réduit (Modèle P) est adéquat, cette statistique est distribuée selon une loi de χ^2 avec pour degré de liberté la différence entre le nombre de paramètres du Modèle G et celui du Modèle P ($n_G - n_P$). À partir de la statistique D obtenue, un test d'hypothèse est effectué : la valeur de D est comparée à la distribution théorique que cette statistique devrait avoir selon l'hypothèse nulle pour voir si la probabilité que le D obtenu appartienne à cette distribution est inférieure à 0.05. Pour illustrer la comparaison de modèles, nous construisons un nouveau modèle qui contient les mêmes variables que le Modèle 3, avec en plus, la variable `Word` comme effet aléatoire. Cette variable spécifie quel est le mot présenté aux sujets. Nous observons, sur la figure 2.14, que les moyennes de temps de décision lexicale varient d'un mot à l'autre. Autrement dit, la variable `Word` forme des groupes dans les données. C'est ce qui justifie sa prise en compte en tant qu'effet aléatoire. Cependant, à la différence de la

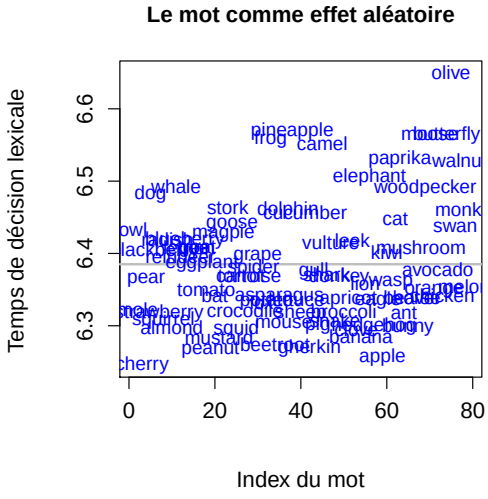

FIGURE 2.14.: Moyenne du temps de décision lexicale pour chaque mot de l'expérience (*lexdec*). La droite grisée indique la moyenne pour l'ensemble des données.

variable `Subject`, l'effet aléatoire `Word` n'a qu'un intercept aléatoire, et pas de pente aléatoire. Le modèle se présente comme dans la figure 2.15. Nous savons

$$RT_{\mathbf{subject=i,word=j}} = \alpha + \alpha_i + \alpha_{\mathbf{word=j}} + (\beta + \beta_i) \times Frequency + \epsilon$$

FIGURE 2.15.: Modèle 4

que $log(L_{Modele3}) = 451.22$ et que $log(L_{Modele4}) = 485.15$. Nous obtenons donc : $D = 2(log(L_{Modele4}) - log(L_{Modele3})) = 2(485.15 - 451.22) = 67.86$; et pour le degré de liberté : $dl = n_{Modele4} - n_{Modele4} = 7 - 6 = 1$. Or pour $\chi^2 = 67.86$ avec un degré de liberté égal à 1, p< 0.05. La statistique D obtenue nous permet de rejeter l'hypothèse nulle selon laquelle le Modèle 3 est le modèle adéquat. L'ajout d'un effet aléatoire `Word` semble donc être justifié pour rendre compte des données *lexdec*.

Nous avons présenté les modèles à effets aléatoires qui permettent de modéliser une variable en fonction de n variables prédictrices, aussi appelées effets fixes, tout en tenant compte de la structuration des données autour des variables à effets aléatoires. Nous avons illustré notre propos par un cas contenant un intercept aléatoire ainsi qu'une pente aléatoire. Dans la suite de ce travail, nous parlerons principalement d'un cas plus simple, où la variable aléatoire présente seulement un intercept aléatoire, comme pour la variable `Word` dans le Modèle 4.

2.2.2. Régression logistique

Dans cette section, nous abordons les modèles de régression logistique qui constituent l'outil principal de modélisation de cette thèse. Ces modèles sont plus complexes que les modèles linéaires, mais la méthodologie qui leur est associée reste la même que celle que nous avons vue précédemment. Après avoir construit un modèle logistique à partir des données de l'échantillon, il faut l'évaluer en statuant sur la significativité des coefficients et en estimant sa capacité de prédiction. Il est également nécessaire de compacter le modèle en sélectionnant, par comparaison de modèles, les variables participant significativement au modèle. Enfin, l'interprétation des modèles logistiques demande le même type de précautions que pour les modèles linéaires (multicolinéarité des variables prédictrices et échelle des variables). Nous présenterons, comme nous l'avons fait pour les modèles linéaires, trois modèles par ordre croissant de complexité : la régression logistique simple, la régression logistique multiple et la régression logistique à effets mixtes.

L'idée de la régression logistique est de modéliser le comportement d'une variable

nominale [28], à la différence de la régression linéaire qui permet de modéliser le comportement d'une variable continue. Une variable nominale a pour valeur des catégories. En linguistique, on utilise souvent ce type de variables :

- le nombre : *singulier*, *pluriel* ;
- le cas : *nominatif*, *accusatif*, *génitif*, *datif* …
- la catégorie lexicale : *nom*, *verbe*, *adjectif*, *adverbe* …

Nous nous intéressons à un cas particulier de variable nominale : la variable binaire. Il s'agit d'une variable nominale qui présente deux valeurs. Nous introduisons donc la méthode de la régression logistique pour les variables binaires uniquement. Les variables binaires sont généralement codées sous forme numérique. Ainsi, les valeurs numériques 0 et 1 sont associées aux valeurs d'une variable binaire. Par exemple, dans le cas de la variable nombre N, nous pourrions avoir $N = 0$ si le nombre est singulier et $N = 1$ si le nombre est pluriel (ou inversement). Le cas où $N = 1$ est appelé le succès, et le cas où $N = 0$, l'échec. Modéliser le comportement d'une variable binaire revient à prédire la probabilité de succès en fonction de variables prédictrices. Une fois cette probabilité connue, la probabilité d'échec peut être déduite, car pour une variable binaire : $P(N = 0) = 1 - P(N = 1)$.

2.2.2.1. Régression logistique simple

Nous présentons d'abord le cas le plus simple de la régression logistique qui permet de modéliser le comportement d'une variable binaire en fonction d'une seule variable prédictrice.

Pour illustrer notre propos, nous utilisons les données *dative*. Ces données reprennent en partie celles utilisées dans Bresnan *et al.* (2007) et Bresnan & Ford (2010). Elles contiennent des données relatives à l'alternance dative en anglais (cf. exemple (16) du chapitre 1 reproduit en (1)).

(1) a. construction à SP datif : *He gives [the picture]$_{theme}$ [to Mary]$_{dest}$*
b. construction à double objet : *He gives [Mary]$_{dest}$ [the picture]$_{theme}$*

Les données *dative* comprennent le type de réalisation du destinataire comme un SN (construction à double objet) ou un SP (construction à SP datif), pour 3263 cas d'alternance dative. La réalisation du destinataire est représentée par la variable `RealizationOfRecipient` qui présente deux valeurs, *NP* ou *PP*, correspondant respectivement aux valeurs numériques 0 et 1. Pour chacune des réalisations, la table de données présente un ensemble d'informations matérialisées à l'aide de 13 variables. Un extrait de cette table de données est présenté dans la table 2.9 [29].

Revenons à la modélisation de la probabilité qu'une variable binaire ait la valeur 1. Dans le cas des données *dative*, nous cherchons à décrire la probabilité que la variable `RealizationOfRecipient` ait la valeur `PP`, ce que nous notons $P(\texttt{ROR} = 1)$. Nous souhaitons estimer cette probabilité en fonction de la variable prédictrice

28. Une variable nominale est une variable discrète non-ordonnée.
29. En raison de sa taille, cette table se trouve à la fin du chapitre.

`RelativeLength` qui correspond à la différence entre le logarithme de la longueur du destinataire et celui de la longueur du thème [30] :

$$\texttt{RelativeLength} = log(\texttt{LengthOfRecipient}) - log(\texttt{LengthOfTheme})$$

Quand `RelativeLength` $= 0$, le destinataire et le thème ont la même longueur, quand `RelativeLength` > 0, le destinataire est plus long que le thème et lorsque `RelativeLength` < 0, le destinataire est plus court que le thème. Nous nous attendons à ce que la probabilité de succès augmente lorsque `RelativeLength` est positif. Pour illustrer le problème, nous faisons le graphique des proportions de succès de la variable `RealizationOfRecipient` en fonction de chaque valeur de `RelativeLength`. Cela est représenté par les nuages de points dans la figure 2.16. Nous pourrions envi-

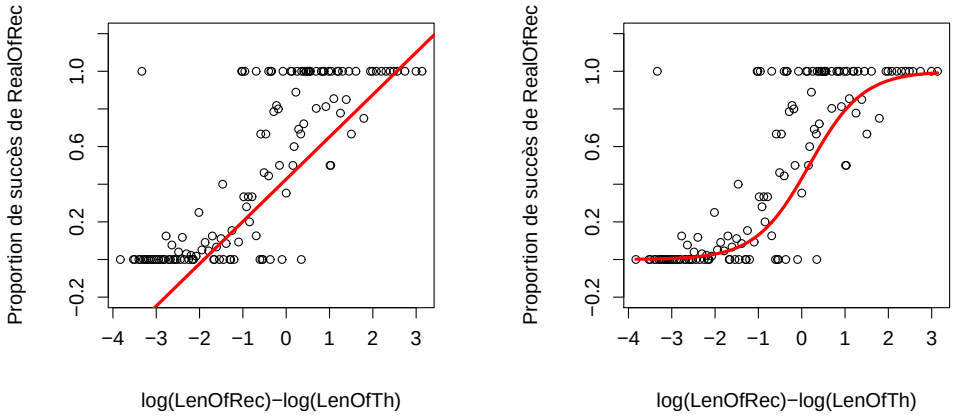

FIGURE 2.16.: Nuage de points représentant la proportion de succès de `RealizationOfRecipient` en fonction de la longueur relative du destinataire et du thème (échelle logarithmique). À gauche, la droite rouge représente la droite de régression linéaire pour ces données. À droite, la courbe rouge est la courbe en forme de S qui est la mieux ajustée aux données.

sager de décrire la probabilité $P(\texttt{ROR} = 1|\texttt{RelativeLength})$ selon un modèle linéaire :

$$P(\texttt{ROR} = 1|\texttt{RelativeLength}) = \alpha + \texttt{RelativeLength}\beta \tag{2.12}$$

Dans ce cas, nous émettons l'hypothèse que la probabilité de succès change linéairement avec `RelativeLength`. Un tel modèle se représente grâce à une droite de régression, comme cela est présenté dans le graphique de gauche de la figure 2.16. Cependant, un tel modèle présente deux inconvénients principaux (Agresti, 2007, p.

30. Nous utilisons les logarithmes des longueurs en nombre de mots, car les variables longueur du destinataire et longueur du thème ont une distribution d'allure exponentielle.

68-70). Premièrement, ce modèle permet de prédire des probabilités supérieures à 1 et inférieures à 0, alors que les probabilités doivent être comprises dans l'intervalle [0,1]. Nous observons sur la figure 2.16 que la droite de régression n'est pas bornée entre 0 et 1. Deuxièmement, la relation qui unit la probabilité de succès et la variable `RelativeLength` semble plutôt non linéaire. En effet, un changement au niveau de la variable `RelativeLength` semble avoir moins d'impact quand la probabilité de succès est proche de 0 ou de 1, que quand celle-ci est autour de 0.5. La relation entre la probabilité de succès et la longueur relative est mieux représentée par une courbe en S. Une telle courbe est présentée sur le nuage de points de droite dans la figure 2.16. La fonction mathématique qui a cette allure en S est la fonction de régression logistique. Elle est définie par la formule suivante :

$$P(Y = 1|x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{2.13}$$

Dans cette fonction, le coefficient α détermine la translation de la courbe, le paramètre β en détermine l'incurvation et la direction. Les propriétés des paramètres α et β sont illustrées dans la figure 2.17. Lorsque α diminue, la courbe se "déplace" vers la droite. Lorsque β est positif, la courbe est en S, tandis que quand il est négatif la courbe suit un S à l'envers. Enfin, plus β est élevé, plus la courbe est incurvée. Le problème est alors d'estimer les paramètres α et β dans une fonction

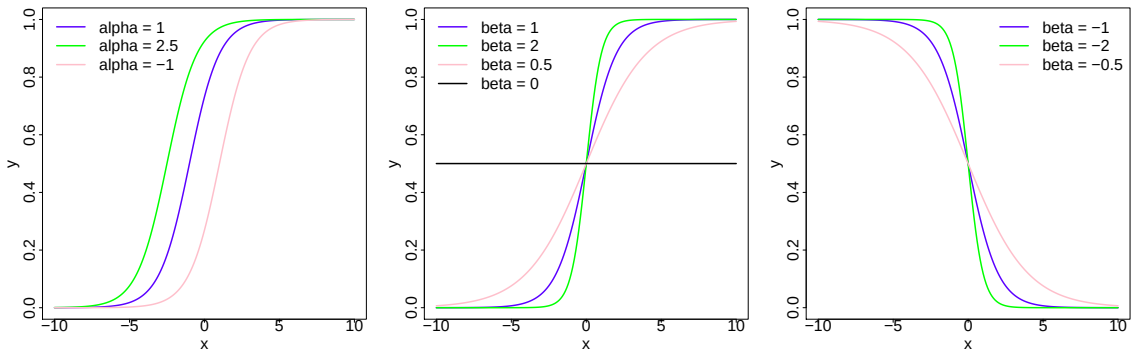

FIGURE 2.17.: Allure de la courbe définie par la fonction logistique selon les valeurs de α et β.

de ce type. Nous pouvons remarquer que la fonction logistique est composée d'une équation linéaire que nous connaissons déjà $(\alpha + \beta x)$ et pour laquelle nous savons estimer les coefficients. L'idée est d'utiliser les méthodes déjà connues. Pour cela, il faut projeter les données qui nous intéressent dans un espace linéaire, estimer les paramètres de régression dans ledit espace et convertir les paramètres estimés dans l'espace logistique, qui n'est pas un espace linéaire. Cette idée est illustrée par les trois graphiques présentés dans la figure 2.18. Dans le premier graphique, nous observons le nuage de points correspondant à notre problème (proportion de succès en fonction de `RelativeLength`). Dans le second graphique, les données sont projetées dans un

espace linéaire grâce à la fonction *logit* que nous décrirons ci-dessous. Dans cet espace linéaire, la droite de régression la mieux ajustée aux données peut être calculée. Cette droite de régression est elle-même projetée dans l'espace logistique, comme cela est montré sur le troisième graphique. D'un point de vue formel, la fonction qui permet

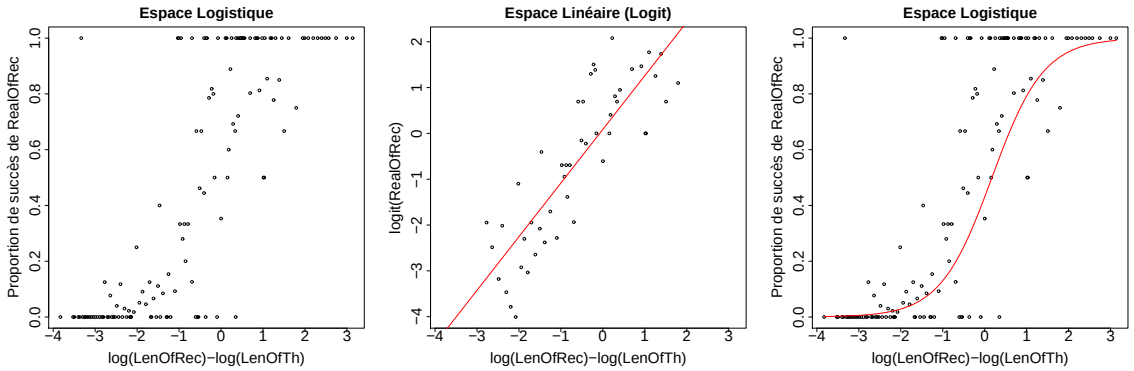

FIGURE 2.18.: À gauche : nuage de points de la proportion de succès en fonction de `RelativeLength`. Au milieu : nuage de points projeté dans un espace linéaire avec sa droite de régression. À droite : courbe de régression qui est image de la droite par $\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$.

de projeter des données logistiques dans un espace linéaire est la fonction *logit*. Cette fonction se définit de la façon suivante :

$$logit(x) = log\left(\frac{x}{1-x}\right) \tag{2.14}$$

Il est possible de montrer que : $logit(P(Y = 1|x)) = logit(\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}) = \alpha + \beta x$. L'objectif est d'estimer α et β dans l'espace logit. Pour cela, la méthode utilisée est celle que nous avons évoquée pour les modèles à effets mixtes : l'estimation du maximum de vraisemblance (ML). À la différence de la régression linéaire et de la régression à effets mixtes, on n'émet pas l'hypothèse que les résidus sont normalement distribués [31]. Pour les données *dative*, le modèle obtenu est présenté dans la figure 2.19 Nous pouvons, par exemple, calculer la probabilité de succès quand la variable `RelativeLength` est égale à 0.

$$P(\texttt{ROR} = 1|\texttt{RelativeLength}) = \frac{e^{-0.28+1.62\times 0}}{1+e^{-0.28+1.62\times 0}} = \frac{e^{-0.28}}{1+e^{-0.28}} = 0.43$$

Ainsi, d'après le modèle que nous avons construit sur les données *dative*, la probabilité d'avoir une construction à SP datif lorsque le destinataire et le thème sont de même longueur est de 0.43.

31. Notons que, de façon générale, l'approche adoptée ici, qui consiste à modéliser les données relatives au langage en utilisant la régression logistique, est *ad hoc*, dans la mesure où le choix du type de modélisation se fait dans un objectif d'ajustement par rapport aux données. Une justification théorique solide de ce choix serait requise pour asseoir cette démarche. Cette question théorique reste ouverte.

$$P(\texttt{ROR} = 1|\texttt{RelativeLength}) = \frac{e^{-0.28+1.62\texttt{RelativeLength}}}{1 + e^{-0.28+1.62\texttt{RelativeLength}}}$$

FIGURE 2.19.: Modèle 5

2.2.2.2. Régression logistique multiple

De la même façon que pour la régression linéaire, le cas simple de la régression logistique à une variable prédictrice peut être étendu à plusieurs variables prédictrices, continues ou nominales. Le modèle de régression logistique multiple se définit de la façon suivante :

$$P(Y = 1|X) = \frac{e^{\beta X}}{1 + e^{\beta X}} \tag{2.15}$$

où

- Y correspond à la variable binaire à prédire,
- X renvoie au vecteur contenant l'ensemble des variables prédictrices (x_1, x_2 ... x_n)
- β renvoie au vecteur contenant l'ensemble des paramètres du modèle (α, β_1, β_2 ...β_n).

Dans le cas des données *dative*, nous modélisons la probabilité de succès de la variable `RealizationOfRecipient` en fonction de `RelativeLength` et de `AnimacyOfRecipient`. Le modèle obtenu est présenté dans la figure 2.20.

$$P(\texttt{ROR} = 1|X) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$
$$\text{avec } \beta X = -0.3664 + 1.6075\ \texttt{RelativeLength} + 0.9758\ (\texttt{AnimacyOfRecipient} = \texttt{inanimate})$$

FIGURE 2.20.: Modèle 6

2.2.2.2.1. Evaluation de modèle Le modèle construit sur les données de l'échantillon doit être évalué. Nous suivons la même procédure méthodologique que pour les modèles linéaires. Nous présentons deux moyens d'évaluation : le test d'hypothèse qui permet de statuer sur la significativité des coefficients et les méthodes graphiques et numériques qui aident à évaluer la qualité des prédictions du modèle.

Test d'hypothèse sur les coefficients Pour évaluer un modèle logistique, il est nécessaire de savoir si les coefficients associés aux variables prédictrices sont significativement différents de 0, afin de déterminer si la relation observée entre la variable prédictrice affectée au coefficient et la variable à prédire peut être généralisée au-delà de l'échantillon.

Le principe méthodologique reste le même que pour la régression linéaire. Cependant, le test utilisé n'est plus le test t de Student mais le test de Wald [32]. Nous cherchons à montrer qu'un coefficient β estimé sur l'échantillon est significativement différent de 0. Nous posons donc l'hypothèse nulle selon laquelle l'échantillon dans lequel β a été calculé est extrait d'une population où β_{pop} est égal à 0. Cela revient à poser que la probabilité de succès de la variable à prédire est indépendante de la variable prédictrice dans la population. Nous allons évaluer cette hypothèse à l'aide du test de Wald. Dans le cas d'un coefficient régression, la statistique Wald se calcule de la façon suivante :

$$W = \frac{(\beta - \beta_{pop})^2}{\sigma^2(\beta)} \tag{2.16}$$

où β est le coefficient de régression calculé dans l'échantillon, β_{pop} le coefficient de régression de la population et $\sigma^2(\beta)$ la variance de β. Si l'hypothèse nulle est vraie, la statistique de Wald se distribue selon une loi de χ^2 à 1 degré de liberté.

Nous appliquons ce test aux coefficients de régression du Modèle 6. Nous obtenons que :

- pour β_1 associé à `RelativeLength` : $W = 172.9$, ($p < 0.001$) ;
- pour β_2 associé à `AnimacyOfRecipient` : $W = 71.1$, ($p < 0.001$).

D'après le test de Wald, nous rejetons l'hypothèse nulle pour les deux coefficients du Modèle 6. Nous concluons que l'échantillon étudié est extrait d'une population où les coefficients β_1 et β_2 sont différents de 0 : β_1 et β_2 sont significativement différents de 0. Autrement dit, la probabilité de succès de la variable `ROR` est dépendante des variables `RelativeLength` et `AnimacyOfRecipient`.

Qualité de prédiction L'évaluation de la capacité de prédiction des modèles logistiques ne peut pas se faire avec le coefficient de détermination multiple R^2 car une telle mesure repose sur l'hypothèse que les résidus sont distribués de façon normale. On utilise donc d'autres méthodes. Nous utiliserons l'**exactitude**, la **représentation graphique de la corrélation entre données observées et données prédites** et l'**aire sous la courbe ROC**.

Un modèle de régression logistique permet d'estimer la probabilité de succès d'une variable binaire. Afin de prédire une valeur binaire, il faut associer à la probabilité prédite une procédure de décision. Cela revient à fixer un seuil θ au-delà duquel il

32. Pour une présentation du test de Wald, voir le chapitre 1 de Agresti (2007). Notons que certains auteurs, notamment Agresti, estiment que pour de petits échantillons, le test du rapport de vraisemblance est plus fiable que le test de Wald. Cela dit, les données de corpus que nous utilisons ne constituent pas de petits échantillons.

est décidé que le succès est prédit. Par exemple, le seuil peut être fixé à 0.5 : $\theta = 0.5$. Cela signifie que si $P(Y = 1) > 0.5$, $Y = 1$, sinon $Y = 0$. Appliqué au Modèle 6, cela signifie que si $P(\texttt{ROR} = 1) > 0.5$, la valeur `PP` de `RealizationOfRecipient` est prédite, sinon c'est la valeur `NP`. Autrement dit, lorsque $P(\texttt{ROR} = 1) > 0.5$, c'est la construction à SP datif qui est prédite, sinon, c'est la construction à double objet.

À présent que nous avons des valeurs prédites pour la variable `RealizationOfRecipient`, nous pouvons les comparer aux valeurs observées, comme nous l'avions vu dans la partie 2.2.1.1.1. Pour cela, nous utilisons une matrice de confusion semblable à celle présentée dans la table 2.3 et dans laquelle apparaissent les succès et les échecs observés en fonction des succès et des échecs prédits. Les prédictions correctes correspondent aux vrais positifs (VP) et aux vrais négatifs (VN). Pour calculer l'exactitude E du modèle, il faut calculer la proportion de prédictions correctes en fonction du nombre de données :

$$E = \frac{VP + VN}{N} \tag{2.17}$$

Pour le Modèle 6, nous obtenons la matrice de confusion présentée dans le tableau

		Prédits	
		Y=1	Y=0
Observés	Y=1	**Corrects** Vrais Positifs	**Incorrects** Faux Positifs
	Y=0	**Incorrects** Faux Négatifs	**Corrects** Vrais Négatifs

Table 2.3.: Matrice de confusion pour un modèle de régression logistique

de droite dans la table 2.4. Le score d'exactitude est de $E = 0.82$. Cela signifie que, dans 82% des cas, le Modèle 6, associé à la procédure de décision, classe correctement la variable `RealizationOfRecipient`. Pour évaluer la qualité de la prédiction, il faut comparer cette exactitude à celle d'un Modèle Nul qui ne contient aucune variable prédictrice et qui prédit systématiquement le succès. Le Modèle Nul a une exactitude de 0.74, ce qui correspond à la proportion de valeur `PP` pour la variable `RealizationOfRecipient` dans les données *dative*. La matrice de confusion de ce modèle est présentée dans le tableau de gauche de la table 2.4. Le Modèle 6 apporte donc une amélioration à la prédiction par rapport au Modèle Nul. Afin d'éviter les effets de surentraînement (cf. partie 2.2.1.2.1), nous procédons à une validation croisée et nous reportons l'exactitude moyenne μ ainsi que son écart type σ. Pour le Modèle 6, avec une validation croisée à 100 passes, nous obtenons une exactitude $\mu = 0.82$ et un écart type $\sigma = 0.064$. Cela signifie que le Modèle 6 n'est pas du tout collé aux données. Il généralise très bien à partir des données.

Le problème posé par la mesure de l'exactitude est que celle-ci repose sur la fixation d'un seuil arbitraire. Avec un tel seuil, la différence entre une probabilité de 0.55 et

		Prédits		%
		Y=1	Y=0	correct
Observés	Y=1	0	849	0%
	Y=0	0	2414	100%
Exactitude				74.0%

		Prédits		%
		Y=1	Y=0	correct
Observés	Y=1	351	498	41.3%
	Y=0	88	2326	96.4%
Exactitude				82.0%

TABLE 2.4.: Matrice de confusion pour le Modèle Nul (à gauche) et pour le Modèle 6 (à droite)

une probabilité de 0.95 n'est pas prise en compte. Dans les deux cas, c'est le même résultat qui va être prédit, à savoir le succès. Pourtant, une probabilité de 0.55 indique que la chance d'avoir un succès est beaucoup plus faible qu'une probabilité de 0.95. Une première solution consiste à représenter graphiquement les probabilités prédites par le modèle en fonction des proportions de succès observées. Pour cela, nous groupons les probabilités prédites selon 10 sous-intervalles égaux sur l'intervalle [0,1]. La probabilité moyenne dans chaque sous-intervalle est comparée aux proportions de succès observées dans le même sous-intervalle. Le graphique de la figure 2.21 représente les probabilités moyennes prédites par le Modèle 6, en fonction des proportions de succès observées. Pour un modèle parfaitement ajusté aux données, l'ensemble des points doit se trouver sur la droite. D'après ce graphique, le Modèle 6 n'est pas très performant au niveau des probabilités centrales [0.4, 0.6], ce qui explique en partie pourquoi le score d'exactitude (0.82) n'est pas très bon. Il existe également une mesure qui permet d'évaluer la qualité du modèle en prenant en compte les différentes probabilités prédites : l'aire sous la courbe ROC (*Area Under the ROC Curve*, **AUC**). Pour comprendre à quoi correspond cette mesure, il faut d'abord définir la courbe ROC [33]. La courbe ROC met en relation le taux de vrais positifs (*True Positive Rate*, TPR) avec le taux de faux positifs (*False Positive Rate*, FPR). Soient VP, VN, FP et FN respectivement le nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs (cf. table 2.3), les deux taux se calculent de la façon suivante :

$$TPR = \frac{VP}{VP + FN}$$
$$FPR = \frac{FP}{VN + FP}$$

Le rapport TPR donne la proportion de succès prédits correctement par rapport au nombre total de succès prédits. Le rapport FPR indique la proportion d'échecs prédits incorrectement par rapport au nombre total d'échecs prédits. Plus le TPR

33. ROC est l'abréviation de l'anglais *Receiver Operating Characteristic*, qui est traduit par caractéristique de fonctionnement du récepteur.

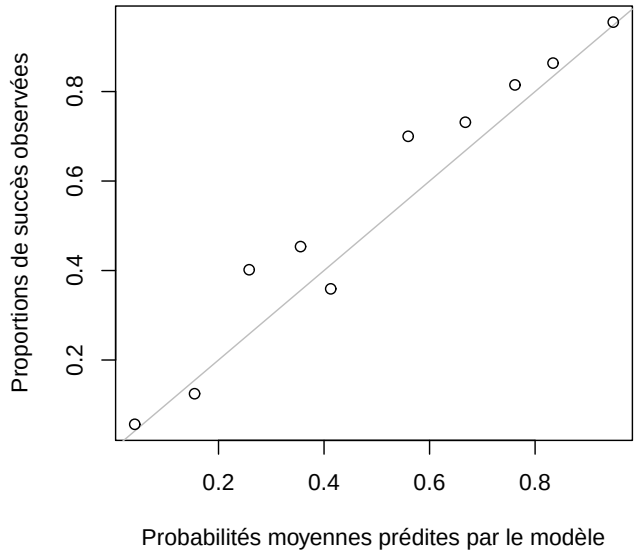

FIGURE 2.21.: Ajustement des observations groupées et des probabilités prédites moyennes pour le Modèle 6

est élevé et plus le FPR est bas, meilleure est la classification. La classification est parfaite lorsque $TPR = 1$ et $FPR = 0$. Pour obtenir une courbe ROC, il faut reporter le TPR en fonction du FPR, pour tous les seuils possibles θ. En d'autres termes, des matrices de confusion sont produites pour tous les seuils possibles, puis les TPR et les FPR sont calculés pour chaque seuil. Chaque point de la courbe renvoie donc à un cas de matrice de confusion. La courbe ROC du Modèle 6 est présentée en 2.22.

Dans le graphique, plus la courbe ROC est proche de la droite grise, plus la capacité du modèle à prédire la classe correcte est mauvaise. Inversement, plus la courbe tend vers le point représentant une classification parfaite, plus la capacité de prédiction du modèle est bonne. D'après la courbe ROC, le Modèle 6 n'a donc pas un très bon pouvoir de classification.

À partir de la courbe ROC, il est possible d'obtenir une mesure de qualité de la prédiction. Cette mesure correspond à l'aire sous la courbe ROC et est symbolisée par le sigle AUC. Plus la courbe tend vers la classification parfaite, plus l'aire sous la courbe est importante. Cette mesure indique la capacité du modèle à discriminer les vrais positifs des faux positifs. Elle peut être interprétée comme la probabilité que le modèle assigne à un exemple positif choisi au hasard une probabilité de succès supérieure à celle qu'il assignera à un exemple négatif choisi au hasard. Une valeur d'AUC égale à 0.5 indique des prédictions aléatoires et une valeur de 1 indique des prédictions parfaites. Généralement, il est admis qu'une valeur supérieure à environ 0.8 a quelque utilité dans la prédiction des ré-

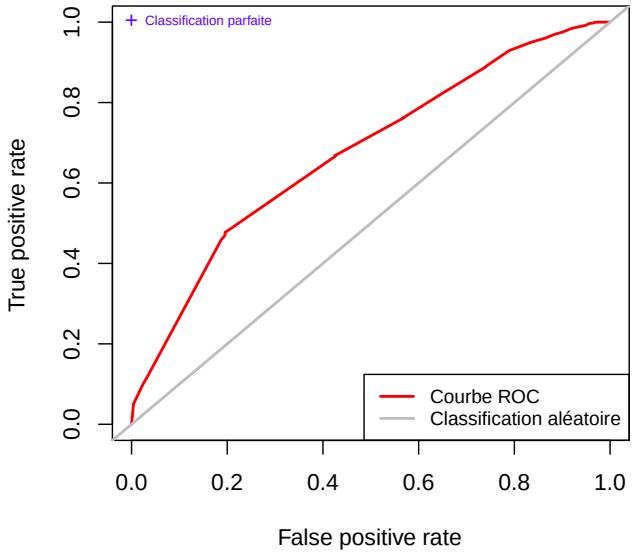

FIGURE 2.22.: Courbe ROC du Modèle 6

ponses (Harrell, 2001, p. 247). Dans le cas du Modèle 6, $AUC = 0.68$. D'après cette mesure, ce modèle ne présente pas vraiment d'intérêt dans la prédiction des réponses.

La régression logistique multiple permet de modéliser une variable binaire en fonction de n variables prédictrices. Comme pour le modèle linéaire, il existe une méthode permettant de trouver le modèle le plus compact, c'est-à-dire le meilleur modèle contenant le moins de variables prédictrices.

2.2.2.2.2. Comparaison de modèles : trouver le modèle le plus compact La comparaison de modèles logistiques se fait grâce au test de rapport de vraisemblance comme pour les modèles à effets mixtes (cf. partie 2.2.1.3). En effet, comme l'estimation des coefficients se fait par maximisation de la vraisemblance des données compte tenu des paramètres du modèle, les vraisemblances des modèles sont comparées afin de vérifier si le modèle le plus compact est suffisant pour expliquer les données.

À titre d'exemple, nous comparons les Modèles 5 et 6. Ces deux modèles sont imbriqués : le Modèle 5 est intégralement contenu dans le modèle 6 et le Modèle 6 présente une variable prédictrice supplémentaire (`AnimacyOfRecipient`). Si nous comparons le Modèle 5 avec le Modèle 6, la statistique D est égale à :

$$D = 2(log(L_{Modele6}) - log(L_{Modele5})) = 2(-1375.316 + 1392.088) = 33.544$$

Nous pouvons rejeter l'hypothèse nulle selon laquelle le Modèle 5 est suffisant pour expliquer les données, car $D = 33.544$ a une p< 0.05. L'ajout de l'effet fixe

`AnimacyOfRecipient` semble donc être justifié pour rendre compte des données *dative*.

2.2.2.2.3. Interprétation des coefficients Pour la régression logistique comme pour la régression linéaire, l'interprétation des coefficients demande que les variables se situent sur la même échelle de valeur et qu'elles ne soient pas (ou peu) corrélées (cf. partie 2.2.1.2.3). De façon générale, un coefficient positif indique que la variable vote pour le succès, tandis qu'un coefficient négatif signale une variable qui favorise l'échec. Cela signifie que dans le Modèle 6, le caractère inanimé du destinataire favorise le succès. Cette interprétation reste vraie tant que le signe des valeurs de la variable est positif. Pour la variable `RelativeLength` dans le Modèle 6, la direction du vote dépend du signe de la variable. Quand `RelativeLength` est négatif, c'est-à-dire que le destinataire est plus court que le thème, la variable favorise l'échec, à savoir la construction à double objet. Inversement, lorsque `RelativeLength` est positif, autrement dit lorsque le destinataire est plus long que le thème, la variable vote pour le succès, soit la construction à SP datif.

L'interprétation des paramètres d'un modèle logistique pose problème. En effet, étant donné que les coefficients β de la régression sont à l'échelle logit, leur interprétation n'est pas intuitive. Cependant, il est possible de convertir ces coefficients en rapports de chance. Pour cela, il faut prendre l'exponentielle du coefficient : e^{coeff}. Par exemple, considérant le coefficient associé à la variable `AnimacyOfRecipient` dans le Modèle 6, nous obtenons le rapport de chance suivant : $e^{0.9758} = 2.65$. Ainsi, d'après le Modèle 6, si le destinataire est inanimé, on a 2.65 fois plus de chances d'observer un succès, à savoir la construction à SP datif.

2.2.2.3. Régression logistique à effets mixtes

Nous avons vu dans la partie 2.2.1.3 que les données sont souvent structurées autour de variables, telles que le sujet ou le lexique. Le même type de groupement des données existe pour les données binaires modélisées grâce à la régression logistique. Pour les données *dative*, nous observons que les différents verbes mis en jeu créent des groupes. Dans le nuage de points de la figure 2.23, nous avons mis en évidence le comportement de trois verbes : *give*, *pay* et *sell*. Nous avons fait apparaître la courbe de régression logistique la mieux ajustée aux données de chaque verbe. Ces courbes font ressortir le fait que les données *dative* sont structurées autour de la variable `Verb`. Il faut donc traiter cette variable comme un effet aléatoire. Les effets aléatoires sont modélisés de la même façon que dans le modèle à effets mixtes (cf. partie 2.2.1.3). Le modèle de régression logistique à effets mixtes se définit de la façon suivante :

$$P(Y = 1|X, Z) = \frac{e^{X\beta+Zb}}{1 + e^{X\beta+Zb}} \tag{2.18}$$

où

- Y correspond à la variable binaire à prédire,
- X renvoie à la matrice représentant l'ensemble des variables prédictrices ;

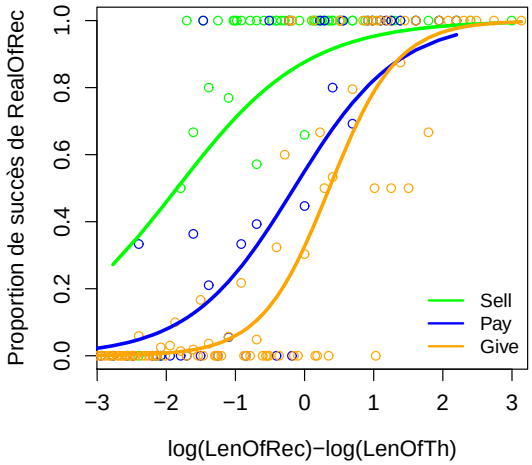

FIGURE 2.23.: Nuage de points représentant la proportion de succès de `RealizationOfRecipient` en fonction de la longueur relative du destinataire et du thème pour trois verbes (*dative*). Les trois courbes de couleur correspondent à la régression pour les verbes *give*, *pay* et *sell*.

- β à la matrice contenant l'ensemble des effets fixes du modèle (intercept et coefficients de pente partiels) ;
- Z renvoie à la matrice représentant l'ensemble des variables à effets aléatoires ;
- b à la matrice contenant l'ensemble des valeurs des effets aléatoires.

Nous voulons maintenant modéliser le comportement de la variable `RealizationOfRecipient` en fonction de deux effets fixes, `RelativeLength` et `AnimacyOfRecipient`, et d'un effet aléatoire `Verb`, qui présente uniquement un intercept aléatoire. Le modèle est donné dans la figure 2.24.

$$P(\texttt{ROR} = 1 | X, b_{\texttt{verb}=i}) = \frac{e^{\beta X + b_{\texttt{verb}=i}}}{1 + e^{\beta X + b_{\texttt{verb}=i}}}$$

$$\begin{aligned} \text{avec} \quad \beta X = \quad & + 0.063 \\ & + 1.64 \ \texttt{RelativeLength} \\ & + 1.21 \ (\texttt{AnimOfRec} = \texttt{inanimate}) \\ \text{et} \quad b_{\texttt{verb}=i} \sim \quad & N(0, 2.08) \end{aligned}$$

FIGURE 2.24.: Modèle 7

Dans la suite de la thèse, les modèles de régression logistique seront présentés sous la forme de tables contenant les paramètres des modèles, ainsi que des statistiques permettant d'évaluer la qualité de la modélisation. Par exemple, la formule de la figure 2.24 se réécrit sous la forme de la table 2.5.

Effets aléatoires :

Groupes	Nom	Variance	Ecart-type
Verb	(Intercept)	4.342	2.0837

Nombre d'obs. : 3263 ; groupes : Verb, 75

Effets fixes :

	Estimation	Erreur-type	valeur z	Pr(>|z|)
(Intercept)	0.06297	0.30627	0.206	0.837
RelativeLength	1.64408	0.07961	20.653	< 2e-16
AnimOfRec = non-animé	1.21105	0.18941	6.394	1.62e-10

Corrélation des effets fixes :

	(Intercept)	RelativeLength
RelativeLength	0.080	
AnimOfRec = non-animé	-0.030	0.060

TABLE 2.5.: Paramètres du modèle 7

En plus des coefficients estimés associés aux variables prédictrices, la table présente l'erreur-type, la valeur z, qui correspond à la statistique de Wald, et la *p-value* associée au test d'hypothèse sur le coefficient estimé (cf. section 2.2.2.3.2). Les effets aléatoires sont accompagnés de leur variance et de leur écart-type. Enfin sont présentées les corrélations des effets fixes qui prennent des valeurs comprises entre -1 et 1 et qui permettent d'estimer la corrélation existant entre des paires d'effets fixes. Plus ces valeurs sont proches de 0, moins la corrélation est importante.

Nous observons que le sens des effets fixes reste le même que dans le Modèle 6, car les coefficients conservent un signe positif. La distribution des valeurs de l'intercept aléatoire (BLUPS) associées à un échantillon de verbes est présentée dans la figure 2.25.

2.2.2.3.1. Comparaison de modèles : trouver le modèle le plus compact Nous utilisons la même méthode de comparaison de modèles que celle utilisée pour le modèle logistique multiple, car l'estimation des coefficients du modèle logistique à effets mixtes se fait également par maximisation de la vraisemblance des données.

Nous comparons le Modèle 6 et le Modèle 7 pour estimer si l'ajout d'un effet aléatoire améliore la qualité de la modélisation. Pour cela, nous pouvons utiliser le rapport de vraisemblance étant donné que les deux modèles sont imbriqués. Le rapport de vraisemblance D pour les Modèles 6 et 7 est le suivant :

$D = -2(log(L_{Modele7}) - log(L_{Modele6})) = 2(-1093.808 + 1733.775) = 1279.934$

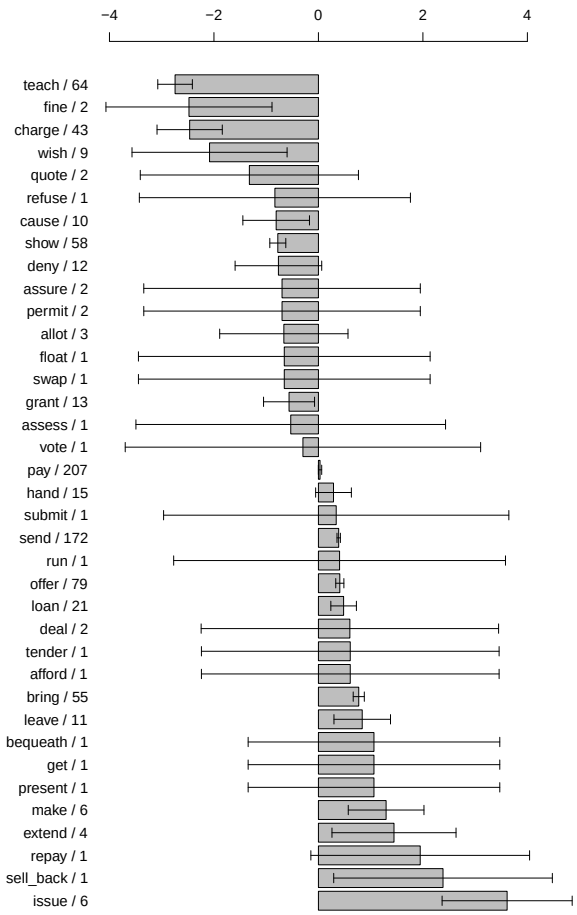

FIGURE 2.25.: Distribution des valeurs de l'effet aléatoire associé à la variable `Verb` pour un échantillon de verbes. Les barres horizontales représentent l'intervalle de confiance à 95% et le nombre accompagnant chaque verbe correspond à la fréquence du lemme dans les données.

Nous appliquons le test d'hypothèse. Nous émettons l'hypothèse nulle selon laquelle le Modèle 6 est suffisant pour modéliser les données. Sous l'hypothèse nulle, la statistique D suit une distribution de χ^2 avec un degré de liberté égal à 1. Or $\chi^2 = 1279.934$ avec un degré de liberté de 1 a une *p-value* inférieure à 0.05 ($p < 0.05$). Nous rejetons donc H_0 et nous concluons que l'ajout d'un effet aléatoire $Verb$ est justifié pour la modélisation des données *dative*.

2.2.2.3.2. Évaluation du modèle Le modèle de régression logistique à effets mixtes est évalué de la même façon que le modèle de régression logistique sans effet aléatoire. Nous présentons, dans cette section, l'évaluation des capacités de prédiction du modèle à effets mixtes que nous avons construit, à savoir le Modèle 7.

Qualité de prédiction Pour évaluer la qualité du Modèle 7, nous construisons la matrice de confusion présentée dans la table 2.6. L'exactitude du modèle est égale à 0.867, ce qui montre que les capacités de prédiction avec un seuil de décision $\theta = 0.5$ sont meilleures pour le Modèle 7 que pour le Modèle 6.

		Prédits		%
		Y=1	Y=0	correct
Observés	Y=1	577	272	70.0%
	Y=0	163	2251	93.2%
Exactitude				86.7%

TABLE 2.6.: Matrice de confusion du Modèle 7

Le graphique de la figure 2.26 représente les probabilités moyennes prédites par le Modèle 7, en fonction des proportions de succès observées. Nous constatons que les points sont bien groupés autour de la droite symbolisant l'ajustement parfait des données prédites aux données observées. Cela indique que la qualité du modèle est bonne. Cette conclusion est confirmée par la valeur d'AUC du Modèle 7 : $AUC = 0.913$. Le modèle a donc une bonne capacité à discriminer les vrais positifs des faux positifs, c'est-à-dire qu'il a une bonne capacité de prédiction pour les données qui nous intéressent.

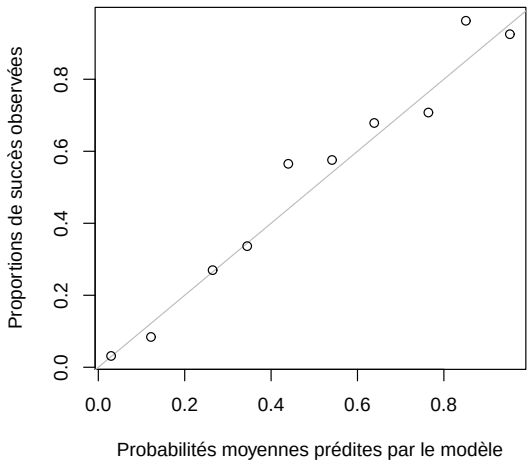

FIGURE 2.26.: Ajustement des observations groupées et des probabilités prédites moyennes pour le Modèle 7.

La régression logistique permet la modélisation d'une variable nominale et, dans

le cas qui nous intéresse, d'une variable binaire, en s'appuyant sur la proportion de succès en fonction des variables prédictrices. La modélisation de la variable binaire revient en fait à la modélisation de la probabilité de succès de cette variable. De la même façon que la régression linéaire, la régression logistique permet de prendre en compte la structure des données grâce aux effets aléatoires. Elle permet également de prédire la probabilité de la variable à prédire pour de nouvelles valeurs des variables prédictrices. Plus cette capacité de prédiction sur des données inconnues est élevée, plus on peut considérer que le modèle a un pouvoir généralisant.

2.2.2.4. Un exemple : la modélisation de l'alternance dative

Nous disposons à présent des outils pour comprendre le modèle proposé par Bresnan *et al.* (2007) et Bresnan & Ford (2010). Les modèles proposés dans ces deux articles sont très semblables. Nous reprenons ici le modèle le plus récent (Bresnan & Ford, 2010) qui présente quelques améliorations par rapport au modèle de Bresnan *et al.* (2007). Etant donné que nous ne disposons pas de la table de données utilisée par ces auteurs, nous reprendrons simplement les éléments décrits dans l'article. Les données exploitées par Bresnan & Ford se composent de 2349 observations de verbes à alternance dative accompagnés d'une construction à double objet ou d'une construction à SP datif, tirées du corpus Switchboard (Godfrey *et al.*, 1992). Le modèle construit à partir de ces données se compose de neuf effets fixes et d'un effet aléatoire. Les variables représentant les effets fixes du modèle sont présentées ci-dessous, avec, entre parenthèses, leurs valeurs possibles.

- `PronOfRec` : caractère pronominal du destinataire (pronominal ou non) ;
- `PronOfThem` : caractère pronominal du thème (pronominal ou non) ;
- `DefinOfRec` : définitude du destinataire (défini ou indéfini) ;
- `DefinOfThem` : définitude du thème (défini ou indéfini) ;
- `AnimOfRec` : caractère animé du destinataire (animé ou non-animé) ;
- `NumbOfThem` : nombre du thème (singulier ou pluriel) ;
- `Previous` : présence d'une construction syntaxique de même type dans le dialogue (construction à SP datif, construction à double objet, aucun) ;
- `RelativeLength` : longueur relative du destinataire et du thème en nombre de mots (`RelativeLength` = log(`LengthOfRecipient`) - log(`LengthOfTheme`)).

Le modèle est présenté dans la table 2.7

L'effet aléatoire renvoie à la concaténation du lemme verbal et de sa classe sémantique[34]. Cet effet aléatoire exprime le biais de chaque verbe dans un emploi spécifique, vers la construction à double objet ou vers la construction à SP datif. Les coefficients des variables à effets fixes doivent s'interpréter de la façon suivante dans le cas des variables binaires (toutes excepté `RelativeLength`) : un coefficient positif indique que la variable vote pour la construction à SP datif, tandis qu'un coefficient négatif accompagne une variable qui penche pour la construction à double objet. En ce qui concerne la variable `RelativeLength`, sa préférence dépend de son propre

34. Les verbes sont annotés selon six classes sémantiques : *abstract*, *transfer of possession*, *future transfer of possession*, *prevention of possession*, *communication*.

Effets aléatoires :

Groupes	Nom	Variance	Ecart-type
verbClassSem	(Intercept)	6.374	2.5246

Nombre d'obs. : 2349 ; groupes : verbClassSem, 55

Effets fixes :

	Estimation	Erreur-type	valeur z	Pr(>|z|)
(Intercept)	1.1583	0.5337	2.170	0.03
PronOfRec = pronominal	-3.3718	0.3236	-10.420	0.000
PronOfThem = pronominal	4.2391	0.4376	9.688	0.0000
DefinOfRec = indéfini	0.5412	0.3147	1.720	0.0001
DefinOfThem = indéfini	-1.5075	0.2877	-5.239	0.000
AnimOfRec = non-animé	1.7397	0.4595	3.787	0.0002
NumbOfThem = pluriel	0.4592	02627	1.748	0.0805
Previous = SP datif	0.5516	0.3406	1.620	0.1053
Previous = aucun	-0.2237	0.2389	-0.936	0.3490
RelativeLength	1.1819	0.1686	7.008	0.0000

TABLE 2.7.: Les paramètres du modèle de Bresnan & Ford (2010)

signe, comme nous l'avons vu pour le Modèle 6 (cf. partie 2.2.2.2.3). Ce modèle est satisfaisant dans la mesure où il présente une faible multicolinéarité et de très bonnes qualités de prédiction, malgré un léger surentraînement. Dans le cas présent, la multicolinéarité est évaluée de façon générale, grâce à l'indice de conditionnement : $c = 8.97$. Ce dernier indique une faible colinéarité des variables prédictrices. La valeur de l'aire sous la courbe ROC, $AUC = 0.984$, indique que la qualité de prédiction du modèle est très bonne. Une évaluation 100 passes montre que le modèle est légèrement surentraîné ($AUC = 0.945$). Pour évaluer l'importance relative des variables prédictrices, les auteurs utilisent la comparaison de modèles. Ils comparent le modèle complet avec un modèle réduit d'une variable en calculant le rapport de vraisemblance. Les variables sont classées selon que leur suppression entraîne une diminution plus ou moins importante de la valeur du rapport de vraisemblance. En d'autres termes, plus la suppression d'une variable fait diminuer le rapport de vraisemblance, plus sa contribution au modèle est considérée importante. Ainsi les variables qui contribuent le plus à la qualité du modèle sont la pronominalité du destinataire et du thème ainsi que la longueur relative. Ce modèle quantitatif montre « *l'existence d'un patron statistique dans lequel, toute chose égale par ailleurs, les arguments animés, définis, pronominaux, accessibles dans le discours et plus courts tendent à précéder les arguments inanimés, indéfinis, non-pronominaux, moins accessibles dans le discours ou plus longs dans les deux constructions datives* » [35].

35. Bresnan & Ford (2010, p. 181) : « *the existence of a statistical pattern in which, all else being equal, animate, definite, pronominal, discourse-accessible, and shorter arguments tend to precede inanimate, indefinite, nonpronominal, less discourse-accessible, or longer arguments in both the*

Les méthodes de régression que nous avons passées en revue permettent de modéliser le comportement d'une variable à prédire, qu'elle soit continue ou nominale, en fonction de variables prédictrices. Associée au test d'hypothèse et aux possibilités de prédiction offertes par les modèles de régression, cette méthode permet de procéder à la généralisation de l'échantillon à la population.

2.3. Expériences psycholinguistiques et études corrélationnelles

L'utilisation des corpus pour étudier des phénomènes de langue offre la possibilité de travailler sur des données "naturelles" en observant les fréquences des unités étudiées, les facteurs intervenant dans le phénomène et les interactions entre les facteurs. Cependant, les données de corpus posent problème pour l'analyse et pour la généralisation. Nous utiliserons deux types d'expérience psycholinguistique pour tenter de dépasser ces difficultés.

Premièrement, nous avons mentionné que la corrélation des variables prédictrices d'un modèle de régression pose problème dans l'interprétation détaillée des modèles et notamment dans l'interprétation du rôle des variables dans le phénomène modélisé. Nous avons évoqué quelques méthodes qui permettent de réduire la multicolinéarité. Cependant, les corrélations sont inhérentes aux données de corpus, dans la mesure où ce sont des données naturelles que l'on ne peut modifier. Cela implique qu'il est quasi impossible d'avoir une corrélation nulle dans un modèle sur corpus. En revanche, l'un des principes fondamentaux de l'expérience psycholinguistique est de décorréler les variables mises en jeu pour pouvoir estimer la valeur explicative de chaque variable sur la variable dépendante. Nous utiliserons donc l'expérience psycholinguistique dans le but de dépasser la corrélation des variables en corpus. L'approche sur corpus et l'approche expérimentale apparaissent donc complémentaires : la première s'appuie sur des données "naturelles" tandis que la seconde repose sur des données "construites" ; l'analyse des données de corpus est limitée par des corrélations que l'on peut contrôler dans les données expérimentales.

Deuxièmement, comme nous l'avons vu précédemment, nous n'avons aucune certitude sur la représentativité du corpus utilisé, car nous ne pouvons qu'émettre des hypothèses sur la diversité et la taille des documents nécessaires à la constitution d'un corpus représentatif. Étant donné que la possibilité de généraliser du corpus vers la langue repose en partie sur la représentativité du corpus étudié, nous ne pouvons qu'émettre des hypothèses sur le niveau de généralisation opérée à partir du corpus. De plus, le corpus est un ensemble d'unités linguistiques sur lequel on applique des méthodes statistiques qui nous permettent de tirer des généralités sur la population de ces unités. Le passage des généralisations sur la population des unités étudiées vers le système de la langue ne va pas de soi. Se pose alors la question de savoir si ce

dative constructions ».

qui a été observé en corpus fait réellement partie de la connaissance des locuteurs. Dans ce cas, des études corrélationnelles seront utilisées dans le but d'apporter des arguments supplémentaires en faveur de l'hypothèse selon laquelle les observations faites à partir du corpus ont une correspondance avec les connaissances des locuteurs.

Les deux types d'expérience envisagés sont très différents, à la fois dans leurs architectures et dans leurs objectifs. Néanmoins, elles présentent un point commun central : elles reposent sur l'élicitation de jugements de locuteurs autour de séquences linguistiques. Ce type d'expérience repose sur une tâche métalinguistique qui a été décrite en détail par Schütze (1996). Nous suivrons les recommandations méthodologiques proposées dans Schütze (1996) et Cowart (1997) pour l'élicitation de jugements, dans le but d'assurer un maximum de fiabilité aux jugements recueillis. Nous tâcherons notamment de limiter les biais dûs à la tâche proposée et ceux introduits par les sujets de l'expérience. Nous reviendrons plus largement sur les précautions méthodologiques lors de la présentation des expériences mises en oeuvre pour les adjectifs et pour les compléments postverbaux.

2.3.1. Élicitation de jugements d'acceptabilité

Cette expérience repose sur des principes généraux de l'expérimentation en psycholinguistique. Il s'agit d'appliquer des méthodes objectives, héritées de la psychologie, au recueil de jugements de locuteurs natifs. La description de la tâche d'un point de vue théorique, et en lien avec la tradition du recueil de jugements de grammaticalité en linguistique, est présentée dans Schütze (1996). Une description plus appliquée des protocoles à suivre est disponible dans Cowart (1997). Schématiquement, ce type d'expérience doit être conçu pour observer l'effet d'une variable étudiée sur la variable dépendante, ici le jugement d'acceptabilité. Soit une variable binaire X ayant les valeurs A et B, le principe de base est de présenter une même phrase dans une condition A et dans une condition B au jugement des locuteurs. L'objectif est d'observer des différences significatives de jugement entre le groupe de phrases de condition A et celui de condition B. Plus précisément, il s'agit d'observer des différences de variances entre les deux groupes. Néanmoins, il existe trois sources de variance (Cowart, 1997, p. 44) qu'il faut contrôler pour permettre de circonscrire au maximum la variance des données et pouvoir l'imputer à la variable X.

La première source de variance est la variance due à la différence entre les groupes A et B ; elle s'appelle variance intergroupe. C'est elle qui nous intéresse dans l'expérience, car c'est elle qui permet de différencier les deux groupes et donc de montrer l'effet de la variable X. Il faut s'assurer que cette variance est bien due à la variable X.

La deuxième source de variance des jugements d'acceptabilité est appelée la variance intragroupe. Elle renvoie à la variance qui existe dans chaque groupe, notamment en raison des différences entre sujets. Cette variance est inévitable, mais il faut en tenir compte dans le plan expérimental et dans l'analyse des données.

Enfin, la troisième source de données est la *variance systématique extérieure*

(*extraneous systematic variance*, Cowart, 1997, p. 45). C'est la part variance, en plus de celle due à la variable X, qui est régulière mais qui ne fait pas partie de la variance manipulée dans l'expérience. Il faut contrôler au maximum cette variance car elle peut être confondue avec la variance due à la variable X. Dans le cas le plus simple, il faut garder constants les autres facteurs influençant (ou supposés influencer) la variable dépendante. Lorsqu'il s'agit d'étudier plus d'une variable, un plan expérimental factoriel est adopté. Ce plan consiste à présenter un croisement systématique de toutes les conditions de chaque facteur. Par exemple, si nous avons deux variables binaires, X et Y, ayant respectivement les valeurs A et B et A' et B', il faut soumettre au jugement des sujets 4 phrases contenant les 4 paires de conditions différentes. Cette idée est présentée dans le tableau de gauche de la table 2.8. Nous reprenons l'exemple de Cowart (1997, p. 48) pour illustrer

		Y	
		A'	B'
X	A	A-A'	A-B'
	B	B-A'	B-B'

		Présence de *that*	
		Sans *that*	Avec *that*
Site d'extraction	Sujet	*Who do you think likes John ?*	*Who do you think that likes John ?*
	Objet	*Who do you think John likes ?*	*Who do you think that John likes ?*

Table 2.8.: À gauche, tableau illustrant le principe du plan expérimental factoriel pour deux variables binaires ; à droite, exemple de plan expérimental factoriel pour les variables "site d'extraction" et "présence de *that*".

le plan expérimental factoriel. Imaginons que l'on veuille tester les variables "site d'extraction" et "présence de *that*" sur les jugements d'acceptabilité en anglais. Le plan expérimental nécessaire est présenté dans le tableau de droite de la table 2.8. Ce type de plan est appelé plan expérimental 2×2 car il est composé de deux facteurs ayant deux niveaux chacun. Utiliser ce type de plan expérimental est un moyen de s'assurer que les variables prédictrices ne présentent aucune corrélation, ce qui permet une analyse statistique exempte des problèmes de colinéarité déjà évoqués. Il est alors possible d'évaluer quelle part de variance attribuer à quelle variable et ainsi estimer l'importance relative de chaque facteur sur le jugement d'acceptabilité. En suivant Baayen *et al.* (2008), nous utiliserons des modèles linéaires à effets mixtes pour l'analyse de ce type de données.

Nous avons exposé rapidement les principes méthodologiques de l'élicitation de jugements d'acceptabilité en mettant en avant l'apport que peut avoir ce type de données par rapport aux données de corpus. Dans la partie qui suit, nous présentons l'étude corrélationnelle que nous mettrons en oeuvre pour tenter de confirmer les résultats obtenus grâce à la modélisation sur corpus.

2.3.2. Préférences sur des paires d'alternatives syntaxiques et corrélation avec un modèle sur corpus

Alors que l'élicitation de jugements d'acceptabilité s'appuie sur des principes méthodologiques classiques et répandus en psycholinguistique, le deuxième type d'expérience que nous appelons études corrélationnelles est moins connu et moins décrit dans la littérature. Nous l'utilisons dans le but de recueillir des jugements de locuteurs et de les comparer avec les probabilités du modèle construit sur corpus. Nous avons rencontré ces études corrélationnelles dans les travaux de Gries (2003b), Bresnan (2007a) et Bresnan & Ford (2010), qui traitent tous les trois de l'alternance dative.

Le plan expérimental repose entièrement sur le modèle construit sur corpus. Le modèle permet de donner la probabilité d'une construction parmi deux alternatives possibles pour une phrase donnée. Cette probabilité peut s'interpréter comme une préférence pour une construction donnée. L'idée de ce type d'expérience est de demander à des locuteurs de procéder de la même façon que le modèle, à savoir, dire quelle est leur construction préférée parmi deux alternatives possibles d'une même construction pour une phrase donnée. Les préférences des locuteurs sont ensuite comparées aux probabilités du modèle, afin de statuer sur la corrélation entre les préférences des locuteurs et le modèle. Si la corrélation est élevée, nous disposons d'un argument en faveur de la qualité de la modélisation, car on peut estimer que les observations faites sur corpus ne sont pas dues à des artefacts et correspondent à une certaine connaissance des locuteurs. Une telle interprétation repose sur l'hypothèse selon laquelle les jugements de préférences sont guidés par des connaissances langagières du même type que celles captées par les variables du modèle sur corpus. Cependant, le lien entre le modèle construit sur corpus et les jugements des locuteurs n'est pas clair. Nous le discuterons à partir des données relatives aux adjectifs et aux compléments verbaux.

La méthodologie de Bresnan (2007b) et de Bresnan & Ford (2010) consiste à présenter le contexte puis la phrase avec les deux alternatives possibles, et à demander au sujet de noter les deux constructions datives selon ses préférences. La méthodologie proposée par Gries (2003b) est sensiblement différente dans la mesure où, pour chaque phrase, une seule alternative (attestée ou construite) est présentée au sujet. Cela signifie que les jugements portés sur la construction dative sont élaborés de façon isolée et non en opposition à l'autre construction possible.

Pour illustrer le type de études corrélationnelles qui nous intéresse, nous reprenons celle proposée par Bresnan (2007b). Afin de montrer que les généralisations tirées sur corpus dans Bresnan *et al.* (2007) ont une réalité chez les locuteurs américains, Bresnan (2007b) propose une étude corrélationnelle dont le but est de déterminer si les préférences des locuteurs correspondent aux probabilités du modèle. L'étude corrélationnelle se compose d'un questionnaire construit à partir de phrases extraites du corpus de Bresnan *et al.* (2007). Il s'agit d'un échantillon de 30 items sélectionnés en fonction de leur probabilité, attribuée par le modèle sur corpus, afin que l'éventail des probabilités soit ventilé de manière homogène entre 0 et 1. Pour chaque phrase

testée, le sujet lit le contexte puis la construction attestée et l'alternative construite, comme cela est montré dans la figure 2.27. Il doit noter les deux constructions datives en distribuant 100 points sur les deux alternatives (par ex. 0-100, 33-67, 50-50).

```
  Speaker:
  About twenty-five, twenty-six years ago, my brother-in-law
showed up in my front yard pulling a trailer. And in this
trailer he had a pony, which I didn't know he was bringing.
And so over the weekend I had to go out and find some wood
and put up some kind of a structure to house that pony,
  (1) because he brought the pony to my children.
  (2) because he brought my children the pony.
```

FIGURE 2.27.: Exemple de phrase proposée dans le questionnaire de Bresnan (2007b)

Pour évaluer la correspondance entre les probabilités des phrases dans le modèle et les notes données par les sujets, Bresnan modélise les notes, en utilisant une régression linéaire à effets aléatoires. Le sujet interrogé ainsi que le sens du verbe constituent les effets aléatoires du modèle. Ces derniers permettent de contrôler la variabilité des données selon le verbe, comme dans la modélisation de Bresnan *et al.* (2007), et selon le sujet interrogé, dans le but de tenir compte du fait que les résultats sont groupés de façon différente selon les sujets (par exemple, certains utilisent la totalité des valeurs possibles [0-100], tandis que d'autres emploient des notes plus ramassées autour de 50). L'auteur observe que les effets fixes, qui correspondent aux variables prédictrices dans le modèle sur corpus, sont tous statistiquement significatifs et qu'ils sont en correspondance avec les tendances observées sur corpus [36]. Ces résultats montrent que les jugements des locuteurs peuvent être expliqués en utilisant les mêmes variables que celles utilisées dans la modélisation sur corpus. Cela constitue un argument pour dire que les contraintes mises à jour sur corpus semblent avoir une existence dans la connaissance langagière des locuteurs.

Dans ce chapitre, nous avons présenté les méthodes et les outils déployés dans cette thèse. Les données qui fondent notre travail sont les données de corpus. Nous avons présenté les problèmes de représentativité et de généralisation ainsi que les corpus du français que nous utilisons. Nous avons montré qu'il existe des outils permettant d'exploiter les données de corpus et de proposer une modélisation satisfaisante. L'analyse des données de corpus s'appuie sur des méthodes de statistique inférentielle qui sont rarement rencontrées dans le champ de la syntaxe. Elles permettent d'envisager les phénomènes étudiés au-delà de l'échantillon exploité. Nous avons également décrit deux autres sources de données que nous utiliserons dans le cadre de

36. Une des variables ne favorise pas la même construction dans l'étude corrélationnelle et dans le modèle sur corpus (*caractère animé du destinataire*). L'auteur justifie cette différence par le fait que seuls deux items présentent une valeur *inanimée* pour cette variable et dans des contextes très spécifiques (Bresnan, 2007b, p. 9).

cette thèse : deux types d'expérience psycholinguistique que nous avons désignés sous le nom d'élicitation de jugements d'acceptabilité et d'élicitation de préférences sur des paires d'alternatives syntaxiques. L'apport de ces sources de données a pour but de dépasser les difficultés relatives aux données de corpus et à leur analyse : corrélation des variables prédictrices et généralisation à la langue.

	Modality	Verb	SemanticClass
1	written	feed	t
2	written	give	a
3	written	give	a
4	written	give	a
5	written	offer	c
6	written	give	a
	AnimacyOfRec	DefinOfRec	PronomOfRec
1	animate	definite	pronominal
2	animate	definite	nonpronominal
3	animate	definite	nonpronominal
4	animate	definite	pronominal
5	animate	definite	nonpronominal
6	animate	definite	nonpronominal
	AnimacyOfTheme	DefinOfTheme	PronomOfTheme
1	inanimate	indefinite	nonpronominal
2	inanimate	indefinite	nonpronominal
3	inanimate	definite	nonpronominal
4	inanimate	indefinite	nonpronominal
5	inanimate	definite	nonpronominal
6	inanimate	indefinite	nonpronominal
	AccessOfRec	AccessOfTheme	RealizationOfRecipient
1	given	new	NP
2	given	new	NP
3	given	new	NP
4	given	new	NP
5	given	new	NP
6	given	new	NP
	LengthOfTheme	LengthOfRecipient	
1	14	1	
2	3	2	
3	13	1	
4	5	1	
5	3	2	
6	4	2	

Table 2.9.: Extrait de la table de données *dative*.

Première partie .

Les adjectifs épithètes en français

Chapitre 3

Le problème de la position de l'adjectif épithète – État de l'art

Sommaire

Ce chapitre et le suivant traitent de la position de l'adjectif épithète par rapport au nom. Une large partie de ce travail a été réalisée en collaboration avec Gwendoline Fox et a fait l'objet d'un article publié dans la revue *Lingvisticæ Investigationes* (Thuilier *et al.*, 2012) et de plusieurs articles publiés dans des actes de conférence (Fox & Thuilier, 2010; Thuilier *et al.*, 2010a,b).

En français, l'adjectif épithète peut être antéposé ou postposé au nom, comme dans l'exemple suivant :

(1) a. *une soirée agréable*
b. *une agréable soirée*

La position de l'adjectif épithète par rapport au nom est un phénomène largement étudié en linguistique. Reiner (1968) retrace « *l'histoire de l'étude de la place de l'adjectif épithète en français* » attestant que cette question intéresse les linguistes et les grammairiens depuis au moins le XVI^{e} siècle. Delomier (1980) revient sur les travaux des linguistes durant le XX^{e} siècle, mettant en lumière la diversité des approches utilisées pour aborder le sujet. D'après les travaux traitant de cette question, les facteurs intervenant dans le choix de la position de l'adjectif par rapport au nom sont d'ordres divers : fréquence, longueur, phonologie, morphologie, syntaxe, sémantique, discours, stylistique. Dans ce chapitre, nous apportons des arguments destinés à montrer qu'une majorité de ces contraintes sont préférentielles et donc intéressantes à étudier en utilisant les méthodes décrites dans les chapitres précédents. L'intérêt majeur de la modélisation que nous proposons est de pouvoir étudier différentes contraintes proposées dans la littérature et ainsi observer leur action combinée et leur importance relative.

Ce chapitre est guidé par l'idée générale qu'il existe une possibilité d'alternance de position pour les mots appartenant à la catégorie "adjectif" en français. Cette alternance s'observe sans contrainte pour certains adjectifs. Elle est plus contrainte pour d'autres. En utilisant le corpus constitué par les documents accessibles sur internet et consulté grâce au moteur de recherche Google, nous essaierons de montrer que l'alternance est envisageable pour la plupart des adjectifs. La prise en compte d'exemples attestés montre que les jugements de grammaticalité sur des exemples construits « *sous-estiment l'espace de possibilité grammaticale* » (Bresnan, 2007a, p. 1). Il y a plus d'alternance que ce que l'on admet généralement. Quelques-unes des phrases trouvées sur internet pourront paraître à la limite de l'acceptable pour certains locuteurs natifs. Néanmoins, nous estimons que leur production témoigne de la possibilité d'alternance au niveau de la catégorie : un adjectif attesté dans une position inattendue peut s'analyser comme un élément qui suit la règle générale associée à la catégorie des adjectifs, selon laquelle il existe deux positions possibles pour l'adjectif épithète en français. Cette règle très générale est confrontée à de nombreuses contraintes préférentielles, relevant notamment du lexique. Ces contraintes peuvent favoriser très fortement une position, ce qui rend "dispréférée" la position inverse pour un adjectif donné.

3.1. Position par défaut

Il est généralement admis que la position par défaut de l'adjectif épithète est la postposition. Nous citons deux arguments permettant de soutenir cette affirmation. Premièrement, tout nouvel adjectif est postposé, qu'il s'agisse d'anciens participes, d'adjectifs de relation construits à partir d'un nom, d'emprunts ou de substantifs employés comme épithète (Noailly, 1999, p. 92). Nous prenons ici l'exemple du néologisme *facebookien*, adjectif de relation construit à partir du nom propre désignant le célèbre réseau social.

(2) *Gameblog n'abuserait-il pas de son pouvoir facebookien ?*[1]

Le deuxième argument repose sur des données quantitatives obtenues dans des travaux sur corpus. D'abord, Forsgren (1978) dénombre 67.2% d'adjectifs postposés parmi les 3 748 occurrences que compte son corpus extrait des journaux *Le Monde* et *L'Express*[2]. Ensuite, sur un ensemble de 29 016 occurrences d'adjectifs relevés dans 80 oeuvres littéraires du XX^{e} siècle, Wilmet (1981) relève 66.4% d'adjectifs postposés[3].

La postposition de l'adjectif étant la position par défaut, nous présentons les différentes contraintes qui la renforcent ou qui attirent l'adjectif en antéposition. Nous avons organisé l'exposé de ces contraintes selon différents niveaux linguistiques allant de la phonologie jusqu'au discours.

3.2. Liaison et hiatus

La position de l'adjectif peut être influencée par les phénomènes de liaison qui existent entre le nom et l'adjectif. Plus précisément, dans certains cas, une position a tendance à être privilégiée, car l'autre position peut causer des hésitations en ce qui concerne la liaison. La liaison correspond à l'insertion d'un segment consonantique à la fin d'un mot, lorsque ce dernier est suivi d'un mot ayant une initiale vocalique.

1. http://www.gameblog.fr/blogs/nohiro/p_56665_gameblog-n-abuserait-il-pas-de-son-pouvoir-facebookien, page consultée le 17 mai 2012.

2. Dans son corpus, Forsgren a éliminé les adjectifs relationnels du type *économique*, *français* et *présidentiel*, ainsi que les adjectifs de couleurs, les ordinaux et les participes passés. Les 3 748 occurrences renvoient à des cas où le nom n'est modifié que par un seul adjectif épithète.

3. Un autre argument est souvent avancé pour étayer l'idée selon laquelle la postposition est la position par défaut de l'adjectif épithète (Noailly, 1999, p. 92). Dans des séquences telles que *un angoissé mythomane* ou *un mythomane angoissé*, c'est l'élément postposé qui est analysé comme l'adjectif épithète. Ainsi, lorsque les deux lemmes combinés peuvent avoir chacun le statut de nom ou d'adjectif, le premier est toujours interprété comme le nom, le second comme l'adjectif. Cependant, il semble que cela ne soit pas vrai dans tous les cas. Par exemple, dans la séquence *une jeune inconnue*, le mot *jeune* est très facilement analysable comme un adjectif antéposé au nom *inconnue*. Étant donné que l'adjectif *jeune* a une forte préférence pour l'antéposition (98.8% d'antéposition dans le corpus de Wilmet, 1981) et que *inconnu* est un adjectif plutôt utilisé en postposition (3.2% d'antéposition chez Wilmet, 1981), il semble que l'analyse de ce type de séquence ambiguë soit guidée par la position la plus fréquente de chacun des éléments.

Ce phénomène est conditionné par la structure syntaxique, mais il existe des variations selon les régions [4], les locuteurs, le type de discours [5] etc. Nous n'avons pas étudié la production des locuteurs. Nous reprenons ici quelques éléments généraux tirés notamment de Mallet (2008) et de Bonami & Delais-Roussarie (à paraître).

Le phénomène de liaison est différent selon que l'adjectif est antéposé ou postposé. Dans le cas de l'antéposition, la liaison est possible mais facultative, au singulier – exemple (3) – comme au pluriel – exemple (4). Dans les exemples, le symbole '*(=)*' entre deux mots note que la liaison est facultative et le symbole '=' que la liaison est réalisée.

(3) a. *un charmant(=)endroit*
b. *un important(=)accord*

(4) a. *ces derniers(=)arguments*
b. *les principaux(=)états*

Il existe deux cas particuliers où la liaison est obligatoire : avec l'adjectif *petit* au masculin singulier et avec les adjectifs indéfinis (Mallet, 2008, p. 79).

(5) a. *un petit=animal*
b. *mes quelques=amis intimes*

Dans le cas de la postposition, la liaison est impossible lorsque le syntagme est au singulier [6], alors qu'elle est possible au pluriel. Dans les exemples en (6), la production de la liaison rend la séquence agrammaticale. En revanche, lorsque le syntagme est au pluriel, la production de la forme de liaison est possible même si elle n'est pas obligatoire, comme le montrent les exemples en (7).

(6) a. **le taux=annuel*
b. **le géant=américain*
c. **un rendement=actuel de 8%*

(7) a. *des avions(=)américains*
b. *des programmes(=)intéressants*
c. *des tragédies(=)intolérables*

Après avoir esquissé les éléments généraux concernant la liaison entre le nom et l'adjectif, nous nous concentrons sur un cas particulier de contexte de liaison qui

4. Grevisse & Goosse (2007) notent que « *Les Parisiens ont tendance à abandonner les liaisons qui se maintiennent mieux en province et en Belgique* ».

5. Grevisse & Goosse (2007) soulignent qu'« *on entend beaucoup moins de liaisons dans la conversation ordinaire que dans le discours soigné et la lecture à voix haute* ».

6. L'impossibilité d'avoir une liaison au singulier n'est pas spécifique à la séquence Nom - Adjectif. De façon générale, la liaison entre le nom et ce qui le suit est impossible au singulier :

(i) **Un cas=à étudier*

pourrait influencer le choix de la position de l'adjectif. Ce phénomène, relevé par Morin (1992), concerne la forme de liaison d'adjectifs masculins singuliers (que nous nommerons dorénavant FLMS), tels que *blanc* ou *franc*. Nous reprenons les exemples et les jugements de Bonami & Boyé (2003, 2005).

(8) a. *ma brune amie*
b. *mon brun camarade*
c. **mon brun ami*

(9) a. *une franche discussion*
b. *un franc dialogue*
c. **un franc entretien*

(10) a. *ma sotte amie*
b. *mon sot camarade*
c. **mon sot ami*

(11) a. *une ambiance chaude / une chaude ambiance*
b. *un débat chaud / un chaud débat*
c. *un entretien chaud / *un chaud entretien*

On observe que, dans un contexte de liaison, c'est-à-dire lorsque le nom commence par une voyelle, la forme du masculin singulier pose des difficultés, tandis que, dans tous les autres contextes (formes du féminin ou du pluriel et forme du masculin hors contexte de liaison), l'adjectif antéposé est acceptable. Cela signifie que c'est la FLMS qui pose problème dans le contexte de liaison. Il est difficile d'estimer l'amplitude du phénomène dans la mesure où ces données ne sont *a priori* pas produites. Bonami & Boyé (2005) citent, en plus de *sot*, *franc*, *chaud* et *brun*, les adjectifs *blanc*, *blond*, *froid*.

Dans les travaux traitant de cette question, l'antéposition de ces adjectifs est considérée comme agrammaticale. Par exemple, Bonami & Boyé (2003, 2005) proposent de traiter ces adjectifs comme étant défectifs pour la FLMS. N'ayant pas de forme disponible pour les contextes à liaison tels que ceux présentés dans les exemples (8-c) à (11-c), les séquences ne peuvent pas être produites par la grammaire.

De notre point de vue, il ne s'agit pas de séquences agrammaticales, dans la mesure où on peut en trouver des attestations, telles que celles présentées en (12) et (13).

(12) *Au coeur du long et* ***franc*** *entretien entre eux, la question du dialogue et de la nécessaire décrispation de la scène politique.*[7]

(13) *Renaud Bergonzo nous présente ce jeune et* ***blond*** *artiste qui photographie des banlieues désaffectées.*[8]

Ces exemples ont été produits à l'écrit, ce qui peut marginaliser l'effet du problème posé par la FLMS. De plus, ils contiennent des coordinations d'adjectifs, ce qui peut,

7. http://opinion.ufpweb.org/politique/dialas.htm, page consultée le 15 février 2012.
8. http://www.delphineaparis.com/tags/renaud-bergonzo, page consultée le 15 février 2012.

comme nous le verrons dans la section 3.4.3, favoriser la mobilité des adjectifs. Bien que de telles séquences existent, on peut émettre l'hypothèse qu'elles sont peu fréquentes, ce qui explique, en partie, qu'elles soient jugées agrammaticales dans les travaux cités. Cette basse fréquence pourrait être analysée comme la conséquence du manque de FLMS. En effet, cette défectivité entraîne une hésitation entre la production d'un hiatus et la production de la forme du féminin pour faire la liaison [9].

La postposition de l'adjectif permet d'éviter d'être confronté à cette situation d'hésitation. Nous posons donc l'hypothèse que l'hésitation entre la liaison et le hiatus disqualifie très fréquemment le choix de l'antéposition pour les adjectifs défectifs pour la FLMS. Dans le cadre de notre travail sur la position des adjectifs, la défectivité des adjectifs pour la FLMS est une contrainte préférentielle favorisant la postposition.

On peut envisager un autre cas où l'absence de liaison pourrait avoir un effet sur le choix de la position de l'adjectif. Il s'agit de la possibilité d'utiliser l'alternance de position pour éviter le hiatus transitoire, à savoir la production de deux voyelles dans la succession de deux mots.

Certains auteurs ont postulé l'existence d'une tendance à éviter le hiatus en français. C'est le cas notamment de Steriade (1999) et Tranel (2000) qui émettent l'hypothèse, dans le cadre de la Théorie de l'Optimalité (Prince & Smolensky, 2004), d'une contrainte interdisant notamment que deux voyelles se succèdent à la frontière de deux mots. Cette idée est loin de faire l'unanimité. Ågren (1973) a étudié un corpus de conversations radiophoniques, dans lequel il a relevé les cas où est réalisée la liaison facultative entre un nom et un adjectif postposé dans un syntagme au pluriel. L'auteur observe que la liaison n'est pas réalisée plus souvent quand le nom finit par une voyelle que lorsqu'il finit par une consonne. Cela semble indiquer que, dans le cas des liaisons facultatives, la contrainte anti-hiatus n'est pas vérifiée, même comme une contrainte préférentielle. De plus, Morin (2005) estime que la contrainte interdisant la succession de voyelles entre deux mots est « *un simple moyen mécanique* » (Morin, 2005, p. 16) permettant d'assurer la liaison dans les cas où elle est obligatoire, mais que cette contrainte n'est pas opératoire lorsque la liaison est facultative.

On sait que la liaison entre le nom et l'adjectif postposé est impossible au singulier.

9. Morin (2003) présente les résultats d'une expérience au cours de laquelle il a demandé à des locuteurs du français de produire, entre autres, les séquences suivantes :

(i) a. Nous avons eu un franc entretien.
b. On pouvait voir au loin un blanc amas d'étoiles parfumées. (tournure littéraire/poétique)
c. Donnez-moi un sot ananas. (phrase sémantiquement non plausible)

Les phrases étaient contrôlées pour l'homogénéité sémantique et les effets stylistiques. L'auteur constate que « *None of the subjects spontaneously used a [ch]-liaison, and all refused* un franc ch-entretien *or* un blanc ch-amas *when they were later asked for grammaticality judgements. None of them spontaneously used the [t]-liaison for (i-c), although some volunteered it after some hesitations* ». Ces résultats laissent penser que les locuteurs préfèrent produire un hiatus plutôt que d'utiliser la forme du féminin des adjectifs défectifs pour la FLMS.

Cela signifie que lorsque le nom termine par une voyelle et que l'adjectif commence par une voyelle, un hiatus est obligatoirement produit. S'il existe une tendance à éviter le hiatus en français, on peut supposer que, dans le cas où l'adjectif est relativement mobile, il a tendance à être antéposé pour éviter le hiatus potentiel en postposition. Si cette hypothèse est vérifiée, on tendrait à préférer (14-b) à (14-a).

(14) a. *état actuel*
b. *actuel état*

Il est important de noter que l'effet du manque de FLMS et de la contrainte anti-hiatus sur le choix de la position est assez faible comparé à d'autres contraintes, telles que les contraintes lexicales ou syntaxiques. Nous le verrons grâce à l'analyse des données de corpus au chapitre suivant.

3.3. Aspects lexicaux

Nous regroupons dans cette partie différentes dimensions concernant l'item adjectival au sens large et susceptibles d'influencer sa position : la longueur de l'adjectif, sa fréquence, ses caractéristiques morphologiques ainsi que sa classe lexicale. Nous considérons que l'ensemble de ces informations est stocké dans l'entrée lexicale de l'adjectif.

3.3.1. Longueur

La longueur est une dimension qui intervient de façon générale dans les phénomènes d'ordre des mots et d'alternance de constructions. D'un point de vue typologique, lorsque l'ordre n'est pas défini de façon catégorique par des règles syntaxiques, les langues à ordre Verbe - Objet (VO) ont tendance à organiser les constituants du plus court au plus long ; et inversement, les langues à ordre Objet - Verbe (OV) tendent à organiser les éléments du plus long au plus court (Hawkins, 1994) [10]. Pour le français, la tendance est donc à placer les éléments les plus courts avant les éléments les plus longs. Nous désignons cette tendance sous le nom *court avant long*.

Dans le cas de l'adjectif épithète, il s'agit de l'ordre relatif de deux mots : le nom et l'adjectif. La longueur doit donc être envisagée au niveau du mot, ce que nous ferons en utilisant le nombre de syllabes. Selon le principe *court avant long*, si le nombre de syllabes de l'adjectif est inférieur au nombre de syllabes du nom, l'adjectif a tendance à être antéposé, tandis que lorsque le nom a une longueur inférieure à celle de l'adjectif, c'est la postposition qui est favorisée. Ainsi, ce principe pousserait à postposer l'adjectif *avide* au nom *air*, mais à l'antéposer au nom *hippopotame*.

10. Nous reviendrons plus en détail sur cette idée et sur les explications apportées par Hawkins (1994) dans le chapitre qui traite de l'ordre des constituants postverbaux (chapitre 5).

(15) a. *un air avide*
b. *un avide hippopotame*

Comme le notent Abeillé & Godard (1999), ce principe n'a rien de catégorique et les ordres inverses à ceux présentés en (15) sont tout à fait acceptables. Il s'agirait seulement d'une tendance. Ce type de préférences est plus facilement observable sur corpus. Forsgren (1978, p. 81) constate que l'ordre organisé selon les *masses croissantes* est attesté à 44.7% dans ses données. Cependant, dans 26.1% de ces données, l'adjectif et le nom ont la même longueur. Si l'on élimine cette part des données pour laquelle la comparaison de la longueur du nom et de l'adjectif n'est pas pertinente, 56.3% d'entre elles présentent l'ordre *court avant long*. Ainsi, la tendance n'est que très légère. D'autres études menées sur l'ordre relatif de deux mots ont montré que la longueur relative peut être une contrainte pertinente au niveau de l'ordonnancement des mots. Par exemple, Benor & Levy (2006) ont étudié l'ordre des deux items lexicaux dans des séquences de conjoints de même catégorie en anglais : *A and B*. L'étude de 692 séquences conjointes leur a permis de mettre à jour notamment que l'ordre des conjoints est significativement affecté par la contrainte de longueur relative : le conjoint le plus court a tendance à apparaître en premier.

Nous émettons donc l'hypothèse selon laquelle il existe une contrainte de longueur pour la position de l'adjectif : lorsque l'adjectif est plus long que le nom, il a tendance à être postposé, tandis que lorsqu'il est plus court que le nom, il a tendance à être antéposé. Comme le suggèrent les données de Forsgren (1978), si elle est valable, cette contrainte ne présente qu'une légère préférence pour l'ordre *court avant long*[11].

La longueur est aussi envisagée en termes absolus : indépendamment de la longueur du nom, un adjectif court est plutôt antéposé, tandis qu'un adjectif long est plutôt postposé. Dans leurs études sur corpus, Wilmet (1981) et Forsgren (1978) ont constaté qu'une majorité d'adjectifs monosyllabiques est observée en antéposition et qu'à partir de 3 syllabes, les adjectifs sont très fréquemment postposés. Forsgren (1978) observe que, parmi 3 189 adjectifs apparaissant comme unique composant du SAdj, 72.8% des lemmes monosyllabiques, 51.6% des lemmes bisyllabiques et seule-

11. L'idée que les mots s'ordonnent selon leur longueur relative semble aller à l'encontre du principe de Miller *et al.* (1997) (*principle of phonology-free syntax*) selon lequel « *In the grammar of a natural language, rules of syntax make no reference to phonology* » (Miller *et al.*, 1997, p. 68). D'après ces auteurs, il n'existe pas de contrainte grammaticale faisant référence à des informations phonologiques. Notons que les données que nous présentons ne vont pas à l'encontre de cette affirmation, dans la mesure où elles mettent en lumière des tendances que nous interprétons comme l'expression de contraintes préférentielles. Nous n'affirmons pas qu'il existe des règles catégoriques, affectant la grammaticalité, qui imposent que le nom et l'adjectif s'organisent selon leur longueur respective. Rappelons également que nous nous situons dans la lignée des travaux séparant les règles de linéarisation de celles de constituance. Ainsi, le fait d'introduire au niveau de la linéarisation des contraintes préférentielles faisant référence à la longueur, ne signifie pas que la constituance, qui concerne exclusivement la syntaxe, est affectée par des contraintes relatives à la phonologie. Nous envisageons les contraintes de linéarisation comme des contraintes d'interface (syntaxe et phonologie, syntaxe et structure informationnelle...), ce qui permet d'expliquer qu'elles puissent faire référence à la longueur des items lexicaux ou des constituants.

ment 28.3% des lemmes trisyllabiques apparaissent au moins une fois en antéposition. L'auteur donne également le nombre moyen de syllabes des lemmes rencontrés dans chaque position. Ainsi, les lemmes présents seulement en antéposition ont en moyenne 2.3 syllabes, ceux observés uniquement en postposition 2.8, et ceux apparaissant dans les deux positions, 2.6. Ces données montrent bien l'existence d'un lien entre longueur absolue de l'adjectif et position. Il semble que l'effet de la contrainte de longueur absolue soit beaucoup plus massif que celui de la longueur relative. Cela laisse supposer que la contrainte de longueur relative peut n'être qu'une conséquence de la contrainte de longueur absolue [12].

A partir de ses 29 016 occurrences d'adjectifs, Wilmet (1981) observe que plus les adjectifs de son corpus sont fréquents, plus ils ont tendance à être monosyllabiques : 9 monosyllabiques sur les 10 adjectifs les plus fréquents, 38 monosyllabiques sur les 50 plus fréquents et 58 monosyllabiques sur les 100 plus fréquents. Étant donné que les 100 adjectifs les plus fréquents s'observent en grande majorité en antéposition, on peut voir un lien entre monosyllabicité et antéposition. Cependant, Wilmet privilégie l'effet de la fréquence de l'adjectif à celui de sa longueur.

3.3.2. Fréquence

Il existe une corrélation forte entre fréquence et position : les adjectifs les plus fréquents ont tendance à être antéposés et les adjectifs moins fréquents sont plutôt postposés. Ce lien entre position et fréquence est observé par Wilmet (1980). Dans ses données, la fréquence est celle du lemme dans les données relevées pour l'étude sur la position de l'adjectif épithète. Par exemple, les trois lemmes adjectivaux les plus fréquents sont *grand*, *petit* et *bon*. Ils se présentent en antéposition respectivement à 96.8%, 98.7% et 97.5%. Par contraste, les lemmes qui n'apparaissent qu'une fois chez Wilmet (1980), comme *coléreux*, *invalide* ou *ocre*, sont, dans une large majorité, postposés. Cette tendance n'est pas sans exception. Les adjectifs de couleur, *blanc*, *bleu*, *noir* et *rouge*, font partie des 15 lemmes adjectivaux les plus fréquents dans le corpus de Wilmet et apparaissent pourtant à 97.4% en postposition. De même, l'adjectif *plein* fait partie des vingt adjectifs les plus fréquents de son corpus et il est n'a pas de préférence pour une position ou l'autre (49% de postposition). Wilmet (1981) établit un lien entre la fréquence de l'adjectif et sa distributivité : plus un adjectif est fréquent, plus il est combiné à un grand nombre de lemmes nominaux différents. L'auteur estime qu'un adjectif présentant une distributivité élevée donne l'avantage à l'ordre Adjectif - Nom.

Si ce sont bien les lemmes les plus fréquents qui apparaissent en antéposition, le nombre de lemmes se présentant dans cette position est très inférieur au nombre de ceux qui se présentent en postposition. Wilmet (1981) observe que seuls 175 lemmes adjectivaux sont antéposés dans ses données, soit 4.6% des lemmes. Ainsi, les lemmes rencontrés en antéposition sont très fréquents mais peu nombreux. À l'inverse, les

12. Nous étudierons en détail l'importance des différentes mesures de longueur dans la section 4.2.1 du chapitre 4.

adjectifs postposés sont moins fréquents en moyenne, mais on observe une très grande variété de lemmes.

Une fois le rôle de la fréquence établi, il reste une question : en quoi la fréquence de l'adjectif peut-elle influencer sa position par rapport au nom ? Un élément de réponse peut être apporté par les travaux sur l'effet de la fréquence en diachronie. D'un point de vue diachronique, l'un des effets de la fréquence relevée par Bybee (2009) est le suivant : les mots et les séquences très fréquents sont renforcés dans leur structure morpho-syntaxique et résistent donc aux changements généraux de règles liés à l'évolution de la langue. En ancien français, l'ordre le plus répandu était Adjectif - Nom pour l'ensemble des adjectifs. Cependant, la possibilité de postposer les adjectifs existait déjà. Buridant (2000, p. 211) explique que la postposition pouvait résulter d'une mise en relief ou de considérations stylistiques (assonnance, rime). Sachant que l'ordre prépondérant en ancien français était Adjectif - Nom, on peut émettre l'hypothèse selon laquelle la règle générale, développée en Français Moderne, qui consiste à postposer l'adjectif au nom, a peu affecté les lemmes adjectivaux très fréquents, ces derniers étant résistants aux changements en raison de leur fréquence élevée. Cette idée est en accord avec le constat de Glatigny (1967) : « *Les adjectifs antéposés appartiennent en très grande majorité au fonds ancien de la langue* » (Glatigny, 1967, p. 209).

3.3.3. Morphologie

La morphologie de l'adjectif a un impact sur sa position. De façon générale, les adjectifs morphologiquement construits ont tendance à préférer la postposition. Hormis les converts, les adjectifs construits sont généralement plus longs que les adjectifs simples et se soumettent donc à la contrainte préférentielle de longueur absolue qui favorise leur postposition (cf. section 3.3.1). En plus de ce simple effet de longueur, d'autres caractéristiques des adjectifs construits ont été identifiées comme intervenant dans leur préférence pour la postposition.

D'abord, les adjectifs déverbaux et dénominaux peuvent, pour certains, être substitués par un groupe plus complexe ayant une valeur sémantique équivalente. C'est le cas du déverbal *contestable* qui peut être substitué par la relative *que l'on peut contester* (ou *qui peut être contesté*) (16), ainsi que du dénominal *semestriel* remplaçable par le SP *du semestre* (17) [13].

(16) a. *un système contestable*
b. *un système que l'on peut contester*

(17) a. *le résultat semestriel*
b. *le résultat du semestre*

De même, les adjectifs dérivés de participes présents ou passifs peuvent être analysés

13. Un part importante des dénominaux forme la classe des adjectifs relationnels. Nous reviendrons en détail sur cette classe dans la section 3.8.1.

comme se substituant à une relative. Ainsi, *alarmant* se substitue à *qui alarme* (18) et *attendu* peut commuter avec *qui est attendu* (19).

(18) a. *un niveau alarmant*
b. *un niveau qui alarme*

(19) a. *une décision attendue*
b. *une décision qui est attendue*

La possibilité que présente une certaine catégorie d'adjectifs construits à commuter avec une séquence syntaxiquement plus complexe et obligatoirement postposée, constitue une contrainte préférentielle favorisant la postposition.

Selon Abeillé & Godard (1999), certains participes sont obligatoirement postposés. Les auteurs jugent agrammaticale l'antéposition de *attendu*, *interdit*, *atténuant* et *concurrent*, comme le montrent leurs exemples et leurs jugements reproduits en (20).

(20) a. *une décision attendue / *attendue décision*
b. *des jeux interdits / *interdits jeux*
c. *des circonstances atténuantes / *atténuantes circonstances*
d. *des propositions concurrentes / *concurrentes propositions*

Nous émettons à nouveau l'hypothèse que la position des adjectifs issus de participes, tels que ceux présentés par Abeillé & Godard, n'est pas imposée de façon catégorique, mais suit une contrainte préférentielle qui favorise très fortement la postposition. L'antéposition, bien que largement dispréférée, n'est pas exclue dans des contextes favorables : exemples (21) à (23). Notons que nous n'avons pas trouvé d'exemple avec *concurrent* antéposé sur Google et que, dans l'exemple (23), l'adjectif est modifié par l'adverbe *très*, ce qui facilite sa mobilité [14].

(21) *Je crains de penser qu'à tous, on trouverait de louables intentions, à défaut, d'**atténuantes** circonstances faites de la litanie convenue des lourdeurs de la procédure, de la charge de travail, de la foi du palais, de l'imprévisibilité du geste d'un fou, d'impondérables divers, etc.* [15]

(22) *Des membres de l'Union des magistrats ont insisté pour que l'affaire Malick Noel Seck serve d'exemple aux autres et parvienne à faire retomber la tension autour de l'**attendue** décision des 5 sages en janvier prochain.* [16]

(23) *Ni le très **interdit** saut périlleux arrière-pied de nez de Surya Bonaly (10e), manière de signifier aux juges qu'elle les maudit à jamais.* [17]

14. À propos du rôle de l'adverbe *très*, voir la section 3.4
15. http://www.maitre-eolas.fr/page/18, page consultée le 17 février 2012.
16. http://www.xamle.net/index.php?option=com_content&view=article&id=390:proces-malick-noel-seck-ce-mardi-que-risque-t-il-xamlenet&catid=49:une&Itemid=244, page consultée le 17 janvier 2012.
17. http://www.liberation.fr/sports/0101237092-patinage-les-juges-ont-prefere-la-technique-a-l-artistique-la-puce-lipinski-saute-plus-or-que-kwan, page consultée le 17 janvier 2012.

Cette préférence générale pour la postposition n'est pas valable pour tous les participes : certains participes présents ayant une valeur évaluative peuvent apparaître beaucoup plus facilement en antéposition (24) ; quelques participes passés présentent également la possibilité de s'antéposer, notamment lorsqu'ils sont "intensionnels" [18] (25).

(24) a. *un étonnant résultat / un résultat étonnant*
b. *un charmant garçon / un garçon charmant*

(25) a. *une présumée candidature / une candidature présumée*
b. *un prétendu jugement / un jugement prétendu*

Il existe une classe d'adjectifs morphologiquement complexes qui a un comportement singulier. Ce sont les adjectifs présentant le préfixe privatif *in-/im-/i-*. Ils sont en moyenne assez longs mais ils peuvent être facilement antéposés, comme en témoignent les exemples en (26).

(26) a. *une indispensable réforme*
b. *une infaillible machine*
c. *une irrésistible ascension*

Ils ont même tendance à être antéposés plus fréquemment que l'adjectif dont ils sont dérivés, quand ce dernier existe. Wilmet (1981, p. 31-33) remarque que pour la quasi totalité des paires adjectif préfixé en *in-*/adjectif servant de base à la préfixation, la présence du préfixe accroît le taux d'antéposition. Si nous prenons la paire *suffisant/insuffisant*, l'adjectif *insuffisant* peut être antéposé sans aucune difficulté (27-a), tandis que l'antéposition de *suffisant* semble moins naturelle (27-b).

(27) a. *une insuffisante activité*
b. *une suffisante activité*

Enfin, concernant la morphologie flexionnelle, aucun des travaux que nous avons consultés ne fait mention d'une influence de ce facteur. Pour la majorité des adjectifs, il n'existe pas de variation phonétique entre les formes du masculin et du féminin (Bonami & Boyé, 2005, p. 84). Dans le cas d'adjectifs présentant une alternance de forme, telle que celles présentées en (28), cette alternance ne semble pas avoir une influence directe sur la position de l'adjectif.

(28) a. *un beau journaliste / une belle journaliste*
b. *un journaliste fou / une journaliste folle*
c. *un nouveau ministre / une nouvelle ministre*
d. *un vieux journaliste / une vieille journaliste*
e. *un entretien bref / une entrevue brève*

Seuls les cas de forme de liaison d'adjectifs masculins singuliers évoqués en sec-

18. Pour plus de détails sur les "intensionnels", cf. la section sur les classes lexicales (4.2.4).

tion 3.2 pourraient constituer une contrainte préférentielle, en lien avec le genre, favorisant une position. De plus, le pluriel de l'adjectif n'est marqué qu'en cas de liaison dans la plupart des cas. Seule une partie des adjectifs en *-al* présente une variation de forme. Cependant, cette alternance ne semble pas influencer le choix de leur position, comme le montrent les exemples en (29). Notons que, étant donné que ces adjectifs sont très souvent dénominaux, ils ont tendance à être postposés.

(29) a. *un organisme national / des organismes nationaux*
b. *un axe central / des axes centraux*

3.3.4. Classes lexicales

Il est généralement admis que la position de l'adjectif est influencée par sa classe sémantique. Par exemple, les adjectifs désignant une propriété objective telle que la forme (30-a) ou la couleur (30-b), la catégorie sociale (30-c), administrative (30-d) ou religieuse (30-e), la nationalité (30-f)… apparaissent en postposition.

(30) a. *une table ronde*
b. *une valise rouge*
c. *la condition ouvrière*
d. *une route départementale*
e. *le peuple juif*
f. *l'attaque américaine*

En revanche, les adjectifs "intensionnels" présentés dans les exemples (31) à (34) ont une préférence pour l'antéposition. Ces adjectifs ne dénotent pas une propriété : ils modifient la dénotation du nom ou la manière dont elle est vérifiée par le référent. Par exemple, l'adjectif *vrai* peut être utilisé pour indiquer que le référent vérifie particulièrement bien l'entièreté de la dénotation, comme dans l'exemple (31).

(31) *une vraie catastrophe*

Dans l'exemple (32), l'adjectif *supposé* ne qualifie pas l'individu désigné, mais le degré d'adéquation entre le référent désigné par *communiste* et le sens de ce mot.

(32) *ce supposé communiste*

De même, dans l'exemple (33), l'adjectif *futur* ne porte pas sur l'individu dénoté par *président*, mais sur l'intervalle de temps à partir duquel l'assignation de la propriété *président* est adaptée pour l'individu désigné.

(33) *le futur président*

Enfin, dans l'exemple (34-b), l'adjectif *parfait* postposé au nom assigne la propriété *parfait* à l'ensemble des individus désignés par *fleurs*, tandis que, antéposé comme en (34-a), la propriété de perfection est assignée au degré d'adéquation entre l'objet désigné et les concepts *scélérat* et *imbécile*.

(34) a. *de parfaits imbéciles, de parfaits scélérats*
b. *des fleurs parfaites*

Les adjectifs "subsectifs" ont aussi tendance à être antéposés, notamment les plus fréquents. Leur caractéristique principale est de dénoter une propriété dont l'interprétation varie avec le nom. Dans l'exemple (35), l'adjectif *petit* qualifie la taille de la table et ce relativement aux dimensions standard de l'objet *table*. Si le même adjectif est associé au nom *maison*, l'interprétation de l'adjectif *petit* ne fera pas référence aux mêmes dimensions.

(35) *une petite table*

Enfin, les adjectifs évaluatifs semblent ne pas avoir une préférence pour une position, mais plutôt autoriser une assez grande liberté quant à leur position. On peut notamment remarquer que des adjectifs longs et morphologiquement complexes peuvent apparaître en antéposition sans aucune difficulté, comme le montre l'exemple (36-b).

(36) a. *un film extraordinaire*
b. *un extraordinaire film*

La classe des évaluatifs possède la particularité de permettre les deux positions de façon quasi-équivalente, en neutralisant, semble-t-il, l'effet d'autres contraintes lexicales de longueur absolue, longueur relative ou complexité morphologique.

Malgré des préférences parfois très marquées, les différentes classes d'adjectifs n'ont pas une position unique. Par exemple, les adjectifs de nationalité sont postposés lorsqu'ils renvoient à la nationalité *stricto sensu*. Cependant, lorsqu'ils font référence aux propriétés qui constituent le stéréotype de la nationalité, l'antéposition est possible. Nous reprenons ici l'exemple de Bouchard (1998, p. 142), dans lequel l'adverbe *très* impose une lecture en termes stéréotypiques.

(37) a. *cette invasion très italienne de l'Albanie*
b. *cette très italienne invasion de l'Albanie*

En ce qui concerne les "intensionnels" qui généralement préfèrent l'antéposition, une recherche sur Google permet de voir que les adjectifs antéposés dans les exemples de (32), (33) et (34) peuvent avoir la même interprétation en postposition, exemples (38), (39) et (40).

(38) *je suis toujours surprise de voir le mépris de certains pour des personnes capables de devenir président des Etats-Unis. On a eu la même histoire avec George Bush, l'imbécile* ***supposé****, et certains recommencent avec Obama.* [19]

(39) *Le futsal est un candidat* ***futur*** *pour les Jeux olympiques.* [20]

19. http://streetgeneration.fr/news/actualite/actu/14395/wyclef-jean-ces-artistes-qui-basculent-en-politique/, page consultée le 31 janvier 2012.

20. http://www.usj.edu.lb/sport/new/index.php/sport-de-competition/97-des-universitaires-sur-le-terrain-des-pros, page consultée le 31 janvier 2012.

(40) *Grâce à la bienveillance de tous les gens qui nous dirigent, politiciens et financiers, de l'extrême droite à l'extrême gauche, de l'imbécile* ***parfait*** *au plus grand philosophe, la "Solution Finale" a presque parfaitement réussi.* (in Destin à part de Henry Bily)

En (38), le contexte de la phrase nous amène à interpréter *supposé* comme portant sur la relation entre le nom et son référent. Dans cette phrase, le locuteur ne considère pas que George Bush est un imbécile. Au contraire, il met en doute que le nom *imbécile* soit approprié pour désigner une personne ayant réussi à être président des États-Unis. L'adjectif *futur* apparaît en postposition dans l'exemple (39) et, comme en antéposition, il porte sur l'intervalle de temps à partir duquel l'assignation de la propriété *candidat* sera adaptée pour le référent désigné. Pour l'adjectif *parfait*, en (40), il semble que la séquence *imbécile parfait* ait la même interprétation que celle attribuée au syntagme *de parfaits imbéciles*, à savoir que l'adjectif indique que l'objet désigné a tous les attributs d'un *imbécile*. Il appartient au plus haut degré à l'extension du mot *imbécile*.

De même, les adjectifs subsectifs ont tendance à être antéposés, mais Wilmet (1981, p. 27-28) [21] montre que cette classe sémantique n'a pas une distribution uniforme. D'abord, certains subsectifs préfèrent la postposition à l'antéposition : par exemple, dans les données de Wilmet, *doux* s'observe à 30.7% en antéposition, *chaud* à 15.1%, *rapide* à 18.4% et *fragile* à 14.3%. De plus, les antonymes n'ont pas forcément les mêmes préférences : par exemple, *haut* est antéposé à 76.6%, alors que *bas* l'est à 25.6% ; *faible* est antéposé à 78.9%, tandis que *fort* l'est à 41.8%. Il semble donc que les préférences de la classe des subsectifs soient disparates.

Finalement, il existe bien des classes sémantiques pour lesquelles on observe des tendances quant à la position. Cependant, ces classes ne constituent pas des contraintes dures imposant une position. Nous les concevons comme des contraintes préférentielles qui en interaction avec d'autres contraintes préférentielles guident le choix des locuteurs.

En résumé, les adjectifs ont des préférences lexicales pour une position ou pour l'autre en raison d'une ou plusieurs de leurs caractéristiques lexicales. Nous allons maintenant voir que des contraintes syntaxiques s'ajoutent à ces préférences lexicales.

3.4. Aspects syntaxiques

3.4.1. Dépendant postadjectival

L'unique contrainte catégorique qui impose une position à l'adjectif est une contrainte syntaxique : la présence d'un dépendant postadjectival. Lorsqu'un adjectif épithète est complété (41) ou modifié (42) par un constituant apparaissant à

21. Wilmet (1981) désigne les adjectifs subsectifs sous le nom d'adjectifs *relatifs*.

sa droite, ce dernier doit être obligatoirement postposé (Abeillé & Godard, 1999; Blinkenberg, 1933).

(41) a. *un homme fier de son fils*
b. **un fier de son fils homme*

(42) a. *une femme belle à croquer*
b. **une belle à croquer femme*

3.4.2. Modifieur pré-adjectival

La présence d'un modifieur pré-adjectival est possible en antéposition comme en postposition.

(43) a. *une très agréable soirée*
b. *une soirée très agréable*

Lorsqu'un adverbe accompagne l'adjectif, la longueur du SAdj est plus importante. Or, si jusqu'ici nous n'avons considéré que la longueur de l'adjectif, il semble intéressant d'étudier la longueur du SAdj dans les cas où l'adjectif est modifié. Forsgren (1978, p. 159) observe que, parmi les 559 adjectifs modifiés par un adverbe, 73.4% sont en postposition, alors que seuls 66.1% des adjectifs formant à eux seuls le SAdj apparaissent en postposition. Ces chiffres suggèrent que la présence d'un adverbe favorise la postposition. Forsgren explique ce constat en partie par l'application du principe *court avant long* au couple Nom / SAdj. Le respect de ce principe implique également que plus l'adverbe est long, plus le SAdj aura tendance à être postposé. Blinkenberg (1933, p. 121) note d'ailleurs que la plupart des adverbes en *-ment*, toujours polysyllabiques, imposent la postposition du SAdj. Cependant, la longueur de l'adverbe n'est pas décisive dans le choix d'un ordre. Les adverbes *merveilleusement*, *exceptionnellement* et *extraordinairement*, qui ont tous plus de quatre syllabes, peuvent être antéposés sans aucune difficulté, comme en témoignent les exemples (44), (45) et (46), tirés du web.

(44) *En cette* ***merveilleusement*** *belle journée, j'ai quelques questions pour vous !*[22]

(45) *Dès 1548, à son arrivée à Lyon, nous le voyons en effet prendre part à une* ***exceptionnellement*** *belle et rare édition de luxe du Corpus juris civilis, publiée par les frères Senneton.*[23]

(46) *Je me trouvais à Paris entre deux périodes de vacances, quelques semaines après avoir accouché d'une* ***extraordinairement*** *belle petite fille (3kg 400), et je déprimais.*[24]

22. http://forum.macbidouille.com/lofiversion/index.php/t154736-250.html, page consultée le 21 mars 2012.
23. http://chretienssocietes.revues.org/2726, page consultée le 21 mars 2012.
24. http://www.ciao.fr/Danone_creme_de_yaourt_fraise_fraise_des_bois___Avis_544458,

La longueur du SAdj modifié est donc un facteur favorisant la postposition. Cela implique que plus l'adverbe est long, plus le SAdj a tendance à être postposé.

Abeillé & Godard (1999) soulèvent un autre aspect de l'influence des adverbes : seul un paradigme limité d'adverbe apparaît dans un SAdj antéposé. Ce paradigme est défini comme étant celui des modifieurs de degré, tels que *très*, *trop*, *assez*, *vraiment*, *peu*, *si*. Par contraste, Abeillé & Godard estiment que les adverbes *absolument* ou *véritablement* ne peuvent pas apparaître en antéposition, comme dans l'exemple (47) (Abeillé & Godard, 1999, p. 14).

(47) a. *Une jeune fille véritablement belle*
b. **Une véritablement belle jeune fille*

Cependant, il est possible de trouver des exemples sur Google où ce type d'adverbes accompagne un adjectif antéposé : exemples (48) et (49).

(48) *Je ne voudrais pas dire une bêtise, ni effrayer les coincés (il y en a...), mais il me semble que cette* ***absolument ravissante*** *jeune fille est une trans.*[25]

(49) *Ou je prend le risque de perdre mon code, de le repasser et de trouver une* ***véritablement bonne*** *auto école ?*[26]

Nous émettons l'hypothèse qu'un adjectif modifié par un adverbe placé à sa gauche peut avoir les deux positions. Cependant, selon la nature de l'adverbe, l'antéposition est plus ou moins facile. Plus précisément, les adverbes qui ne sont pas des adverbes de degré favorisent massivement la postposition.

Par ailleurs, les adjectifs présentant une forte préférence lexicale pour une position, voient cette préférence assouplie par la présence d'un adverbe. C'est le cas de l'adjectif *bon* qui présente une forte préférence pour l'antéposition quand il est seul (50-a), mais accepte facilement la postposition une fois modifié (50-b). De même, l'adjectif *familial* préfère très largement la postposition (51-a), mais peut être antéposé quand il est accompagné d'un adverbe (51-b).

(50) a. *un bon poulet / un poulet bon*
b. *un très bon poulet / un poulet très bon*

(51) a. *la berline familiale / la familiale berline*
b. *la berline très familiale / la très familiale berline*

La présence d'un adverbe tel que *très* est donc un facteur facilitant la mobilité des adjectifs ayant des préférences lexicales très marquées.

---

page consultée le 21 mars 2012.

25. http://princessekyonyuu.blogspot.fr/2012/01/toute-de-camel-vetue.html, page consultée le 21 mars 2012.

26. http://forum.doctissimo.fr/viepratique/Permis-de-conduire/besoin-conseils-sujet_2094_1.htm, page consultée le 23 mars 2012.

3.4.3. La coordination

La coordination d'adjectifs est possible en antéposition comme en postposition.

(52) a. *une table belle et longue*
b. *une belle et longue table*

Abeillé & Godard (1999) précisent que seule la coordination simple est autorisée en antéposition. La coordination corrélative est impossible en zone prénominale comme le montre l'exemple (53-b) (Abeillé & Godard, 1999, p. 14).

(53) a. *une table et belle et longue*
b. **une et belle et longue table*

De la même façon que le modifieur pré-adjectival, la coordination simple d'adjectifs assouplit les préférences lexicales. Dans les SN présentés en (54), on observe que l'adjectif *grand* est plus naturel en antéposition, tandis que, pour l'adjectif *calme*, la postposition est plus naturelle. Une fois les deux adjectifs coordonnés, il apparaît que les deux positions sont tout à fait acceptables.

(54) a. *un grand appartement / un appartement grand*
b. *un calme appartement / un appartement calme*
c. *un grand et calme appartement / un appartement grand et calme*

Abeillé & Godard (1999) signalent le cas des intensionnels *vrai* et *faux*, pour lesquels la postposition est difficile quand l'adjectif est seul (55-a,b). L'exemple (55-c) montre que la coordination de deux adjectifs ayant une très forte préférence lexicale pour l'antéposition peut apparaître en postposition.

(55) a. *des faux coupables / des coupables faux*
b. *des vrais coupables / des coupables vrais*
c. *des vrais ou faux coupables / des coupables vrais ou faux*

La coordination est donc un moyen de neutraliser les préférences lexicales.

En ce qui concerne les tendances d'un point de vue quantitatif, on peut poser l'hypothèse que la coordination favorise la postposition en raison de l'importante longueur du SAdj. Forsgren (1978) observe 72.9% de postposition dans le cas des adjectifs coordonnés. Cela indique une légère préférence pour la postposition, étant donné que la proportion d'adjectifs postposés est de 67.2% dans l'intégralité de son corpus.

3.4.4. Autres dépendants du nom

Grevisse & Goosse (2007) évoquent l'existence d'une tendance à équilibrer les éléments à l'intérieur du SN, notamment dans la langue écrite. Par exemple, lorsque le nom est accompagné d'un complément prépositionnel, l'antéposition de l'adjectif permet de ne pas séparer le nom de son complément, comme en (56-c).

(56) a. *un recueil récent / un récent recueil*
b. *un recueil récent de thèmes grecs*
c. *un récent recueil de thèmes grecs*[27]

Plus généralement, la présence d'autres éléments postposés au nom peut favoriser l'antéposition de l'adjectif. Les éléments susceptibles d'apparaître dans le SN sont des relatives (57), des SP (58) ou des adjectifs (59). Dans les SN donnés en (57-c) et (58-c), la présence de la relative ou du SP n'impose pas de position, mais semble rendre l'antéposition de l'adjectif *habituel* un peu plus naturelle.

(57) a. *l'air habituel / l'habituel air*
b. *l'air habituel que joue Paul*
c. *l'habituel air que joue Paul*

(58) a. *sa tête habituelle / son habituelle tête*
b. *sa tête à coucher dehors habituelle*
c. *son habituelle tête à coucher dehors*

Lorsque deux adjectifs sont modifieurs du même nom, comme en (59), la possibilité d'antéposer l'un des deux adjectifs permet d'obtenir un SN plus équilibré.

(59) a. *un animal étrange indomptable*
b. *un étrange animal indomptable*

Nous posons l'hypothèse que la présence d'autres constituants dépendants du nom est une contrainte préférentielle favorisant l'antéposition de l'adjectif. Aucun des travaux sur corpus que nous avons consultés ne fait d'observation concernant ce point.

3.4.5. Déterminant introduisant le SN

Forsgren (1978) émet l'hypothèse que la nature du déterminant est un élément formel permettant de mieux expliquer la position de l'adjectif. Dans ses données, cet auteur observe que l'adjectif a tendance à être postposé quand le déterminant est indéfini (*un*, *de*, *des*, *du*). Il estime que l'antéposition observée avec les indéfinis est principalement due aux adjectifs ayant une valeur évaluative ou quantificatrice.

Lorsque le déterminant est défini (*le*, *les*, *ce*, *ces*, *son*, *ses*), il observe que l'antéposition est légèrement favorisée. La proportion d'antéposition est particulièrement élevée pour les déterminants possessifs. Forsgren considère que ce sont principalement, selon ses propres mots, les valeurs "déictiques" et "modales" de l'adjectif qui expliquent le taux d'antéposition légèrement plus élevé. Par adjectif à valeur "déictique" et à valeur "modale", l'auteur fait référence à ce que nous avons appelé adjectifs intensionnels. Les adjectifs "déictiques" sont ceux qui permettent de localiser temporellement le référent désigné par le nom. Ainsi, sont des adjectifs déictiques

27. http://www.antiquite.ens.fr/IMG/file/pdf_cours_daix/GRECarticlesyntaxe.pdf, page consultée le 22 mars 2012.

les adjectifs *récent*, *présent*, *précédent*, *futur*, *ancien*, *prochain*... Associés à un déterminant défini, ces adjectifs sont souvent antéposés : *le récent sondage*, *le futur président*, *le présent chapitre*. Les adjectifs à valeur "modale" sont les autres adjectifs intensionnels, qui qualifient la relation entre le nom et le référent du nom : *véritable*, *faux*, *éventuel*, *possible*... Dans les données de Forsgren, ces adjectifs s'observent fréquemment en antéposition avec le déterminant défini.

Si la nature du déterminant a une influence sur la position de l'adjectif, il semblerait que les indéfinis favorisent la postposition, alors que les définis préfèrent légèrement l'antéposition. [28]

3.4.6. La fonction du SN

Dans son étude sur corpus, Forsgren (1978) a étudié la position de l'adjectif épithète selon la fonction assumée par le SN dans la phrase. Il a notamment observé l'effet exercé par l'interaction entre la fonction et le type de déterminant, sur la position. Il observe, par exemple, que la fonction sujet est favorable à l'antéposition dans tous les cas, alors que la fonction attribut favorise l'antéposition seulement dans le cas où le déterminant est défini [29]. Pour les autres fonctions étudiées, aucune tendance claire ne se dégage.

3.4.7. Adjectifs dans des constructions à verbe support

Dans les constructions à verbe support, lorsqu'un adverbe, de sens aspectuel ou de manière, modifie le verbe, on constate l'existence d'une construction sémantiquement équivalente où le nom est modifié par l'adjectif morphologiquement lié à l'adverbe, comme dans les exemples (60) à (62).

(60) a. *Luc a pris* ***rapidement*** *une décision*
b. *Luc a pris une décision* ***rapide***

(61) a. *Luc est* ***financièrement*** *sous la tutelle de Marc*
b. *Luc est sous la tutelle* ***financière*** *de Marc*

(62) a. *Marie fait* ***fréquemment*** *des faux pas*
b. *Marie fait des faux pas* ***fréquents***

28. Notons que le déterminant indéfini pluriel connaît une alternance de forme lorsque l'adjectif est antéposé. Alors que ce déterminant est réalisé *des* quand le SN contient un adjectif postposé, il peut être réalisé *des* ou *de* avec un adjectif antéposé (Gross, 1967), comme dans l'exemple (i).

(i) a. *Il a vu* ***des*** *crimes horribles* / **Il a vu* ***de*** *crimes horribles*
b. *Il a vu* ***des*** *horribles crimes* / *Il a vu* ***d'*** *horribles crimes*

Cette variation ne semble pas avoir une influence sur le choix de la position de l'adjectif.

29. Ces observations ont été faites sur un nombre réduit de données. Seuls 57 SN définis ont la fonction attribut dans les données de Forsgren (1978). Cependant, la proportion d'antéposition (43.9%) est plus importante que dans l'ensemble des SN définis (37.5%), ce qui laisse penser que cette fonction peut avoir une préférence pour l'antéposition.

Ce phénomène, désigné sous le nom de “descente de l’adverbe” (Giry-Schneider, 1987, p. 31), n’est pas réservé aux constructions à verbe support. C’est ce que montre les exemples présentés en (63).

(63) a. *Paul a* ***rapidement*** *bu une tasse de thé*
b. *Paul a bu une tasse de thé* ***rapide***

Ce phénomène ne semble pas avoir une influence spécifique sur la position de l’adjectif. En effet, les adjectifs *rapide* et *fréquent* des exemples précédents peuvent apparaître antéposés dans ces constructions, comme en témoignent les exemples (64).

(64) a. *Luc a pris une* ***rapide*** *décision*
b. *Marie fait de* ***fréquents*** *faux pas*
c. *Paul a bu une* ***rapide*** *tasse de thé*

L’adjectif *financier*, pour sa part, est difficilement antéposable dans ce contexte, comme le montre la phrase en (65). Cette difficulté ne relève pas de la “descente de l’adverbe”, mais des caractéristiques lexicales de l’adjectif : en tant qu’adjectif morphologiquement construit et dénotant une propriété objective, sa postposition est largement préférée (cf. section 3.3).

(65) *Luc est sous la* ***financière*** *tutelle de Marc*

La présence d’un modifieur du nom est obligatoire dans certaines constructions à verbe support (Giry-Schneider, 1996). Par exemple, en (66), les modifieurs *sonore*, *que l’on entend de loin* ou *de hyène* doivent obligatoirement être présents pour rendre la phrase acceptable.

(66) a. *Max a un rire* ***sonore***
b. *Max a un rire* ***que l’on entend de loin***
c. *Max a un rire* ***de hyène***
d. *?Max a un rire*

Les phrases présentées dans les exemples (67) à (69) montrent que la position de l’adjectif n’est pas restreinte dans ce type de construction. En effet, l’antéposition comme la postposition sont possibles.

(67) a. *Luc a une* ***grande*** *admiration pour Léa*
b. *Luc a une admiration* ***folle*** *pour Léa*

(68) a. *Luc a une* ***bonne*** *santé*
b. *Luc a une santé* ***fragile***

(69) a. *Luc a de* ***beaux*** *yeux*
b. *Luc a des yeux* ***bizarres***

Ce rapide survol des adjectifs concernés par les constructions à verbe support semble indiquer l'absence de lien entre la position de l'adjectif et ce type de construction [30].

3.5. Aspects sémantiques

Une large part de la littérature relative à la position de l'adjectif épithète est consacrée à la dimension sémantique de cette alternance. De façon générale, la sémantique de l'adjectif ainsi que la sémantique de la combinaison de l'adjectif avec le nom est un problème complexe, comme en témoigne la littérature théorique sur le sujet : Kamp (1975), McNally & Kennedy (2008) parmi beaucoup d'autres. En français, ce problème sémantique entre en interaction avec la position variable de l'adjectif, ce qui donne lieu à un problème d'autant plus complexe qu'il existe bien des liens entre position et sens, mais « *il n'y a pas de propriété sémantique générale qui soit liée de manière parfaitement régulière à l'ordre relatif du N et du A* » (Abeillé & Godard, 1999, p. 12).

L'objectif de cette partie est de montrer que la sémantique ne permet pas de déterminer systématiquement la position de l'adjectif épithète et que les facteurs sémantiques gagnent à être envisagés comme des contraintes préférentielles au niveau de la catégorie générale de l'adjectif.

3.5.1. Les adjectifs homonymes

Pour un petit ensemble d'adjectifs, on admet généralement l'existence de deux homonymes qui se distinguent par leur sens et leur position. Autrement dit, ces adjectifs sont censés avoir un sens systématiquement différent selon leur position. Voici une liste d'adjectifs homonymes telle qu'on la trouve fréquemment : *pauvre*, *pur*, *propre*, *simple*, *brave*, *cher*, *faux*, *sale*, *seul*, *simple*, *vrai*, *sacré*, *ancien*, *commun*. Les exemples suivants, dont les quatre premiers sont tirés de Abeillé (à paraître), illustrent la relation entre sens et position pour ces adjectifs homonymes.

(70) *ce pauvre garçon* vs. *ce garçon pauvre*

(71) *un pur produit* vs. *un produit pur*

(72) *son propre pantalon* vs. *son pantalon propre*

(73) *une simple phrase* vs. *une phrase simple*

(74) *un brave garçon* vs. *un garçon brave*

(75) *une sacrée histoire* vs. *une histoire sacrée*

(76) *un ancien coffre* vs. *un coffre ancien*

30. Étant donné qu'il ne semble pas y avoir de lien entre construction à verbe support et position de l'adjectif, cet aspect ne sera pas étudié à partir des données de corpus dans le chapitre 4.

Pour chaque paire d'exemples, le sens constaté en antéposition est différent de celui observé en postposition d'un point de vue véric onditionnel : un *pauvre garçon* peut être riche ; un *pur produit* peut être *mélangé* ; son *propre pantalon* peut être sale ; une *simple phrase* peut être grammaticalement complexe ; un *brave garçon* peut être un lâche. Cependant, de telles oppositions sont trop radicales dans la mesure où ce n'est pas la position qui définit le sens. En effet, on peut trouver une interprétation normalement réservée à la postposition en antéposition, comme le montrent les exemples suivants.

(77) *Ils habitent tous Avenue Q, le plus* ***pauvre*** *quartier de New-York, à l'opposé des riches Avenues A, B et C.* [31]

(78) *Michel enfila sa plus belle chemise de trappeur à gros carreaux rose et vert et son plus* ***propre*** *pantalon de velours côtelé bleu cyanosé.* [32]

(79) *L'huile de lin est un* ***pur*** *produit naturel qui est pressée à froid sans additif.* [33]

(80) *Pleins de haine ils souhaiteraient se débarasser de Gaza pour prendre la totalité d'Israël, et faire passer cette évacuation (décision unilatérale) pour un* ***brave*** *acte de Paix.* [34]

(81) *Devant un public amoureux d'****anciennes*** *chansons, Patrick a interprété un répertoire très traditionnel.* [35]

(82) *Le fondement juridique est discutable, car il s´attaque au très* ***sacré*** *droit de propriété.* [36]

Notons que, dans certains exemples, l'antéposition est favorisée par la présence d'un adverbe *très* (82) ou *plus* dans la construction *le plus Adj N* (77), (78) et (80). Cela est tout à fait cohérent avec l'idée que la présence d'adverbe constitue une contrainte préférentielle donnant plus de liberté dans le choix de la position de l'adjectif.

L'inverse, c'est-à-dire l'interprétation normalement associée à l'antéposition en postposition, est plus rare, mais il est possible d'en trouver des exemples. L'adjectif intensionnel *ancien*, dans le sens de *ex-*, qui apparaît quasi systématiquement en antéposition, peut se rencontrer en postposition comme en témoigne la phrase (83), extraite de l'ouvrage de Ménager (1979, p. 229).

(83) *L'émergence du mot « folie » correspond à l'inquiétude d'un poète qui cherche encore les moyens de restaurer l'unité* ***ancienne***

31. http://www.evene.fr/theatre/actualite/interview-bruno-gaccio-avenue-q-spectacle-musical-broadway-bobin-780159.php, page consultée le 15 février.

32. http://busterk.blog.lemonde.fr/2010/09/09/la-carte-de-sejour-et-le-territoire-national/, page consultée le 1er février 2012.

33. http://www.vpchorse.com/catalog/huile-de-lin-awa-10-p-1590.html, page consultée le 1er février.

34. http://emmabenji.canalblog.com/archives/2006/04/14/1701136.html, page consultée le 15 février 2012.

35. Extrait du corpus `Est-Républicain` (cf. section 2.1.3, chapitre 2).

36. http://www.jeuxvideo.com/forums/1-27-7666091-1-0-1-0-0.htm, page consultée le 15 février.

Dans ce contexte, la présence du verbe *restaurer* révèle que l'on doit interpréter cette *unité ancienne* comme une *ancienne unité*, qui n'existe plus aujourd'hui. Nous donnons trois exemples supplémentaires, pour les adjectifs *propre*, *pur* et *brave*.

(84) *L'employeur peut également intervenir dans les frais supportés par le travailleur pour se rendre avec sa voiture* ***propre*** *de son domicile à son lieu de travail*[37]

(85) *Peut-on dire que le sarkozysme est le produit le plus* ***pur*** *de la Ve République ?*[38]

(86) *Le personnage de John C. Reilly, un type* ***brave*** *et mou qui se métamorphose au milieu de la scène en gros con macho.*[39]

Il existe bien des adjectifs homonymes qui se rencontrent de façon extrêmement majoritaire dans une position plutôt que dans l'autre. Pour ces homonymes, il faut distinguer deux entrées lexicales. Cependant, nous considérons que le sens de chaque homonyme n'est pas catégoriquement corrélé avec une position.

3.5.2. Position déterminée par la combinaison du nom et de l'adjectif

L'étude de l'interaction entre sémantique et position de l'adjectif doit prendre en compte le nom auquel est associé l'adjectif. Schématiquement, l'idée est que, lorsque l'adjectif est postposé, le référent du nom est complété, enrichi d'une propriété tandis qu'avec un adjectif antéposé, le référent du nom se combine avec la propriété dénotée par l'adjectif pour former une entité nouvelle et unique, « *sentie comme une unité de pensée* »(Grevisse & Goosse, 2007, p.534). Ce contraste sémantique s'illustre bien par la combinaison d'un adjectif avec un nom propre. Nous reprenons ici l'exemple de Noailly (1999, p. 93).

(87) a. *l'Odile mystérieuse*
b. *la mystérieuse Odile*

Dans le cas d'antéposition, la qualification semble inhérente à la personne : la caractéristique *mystérieuse* est une propriété intrinsèque du référent du nom propre *Odile* ; c'est une propriété définitoire. En revanche, lorsque l'adjectif est postposé, l'individu *Odile* se voit attribuer la propriété *mystérieuse* qui sert à identifier cette *Odile* parmi un ensemble d'*Odile* : *l'Odile mystérieuse* s'oppose à un ou plusieurs autres individus nommés *Odile* mais n'ayant pas la propriété *mystérieuse*. Ici, la postposition impose donc une lecture contrastive tandis que l'antéposition ne l'impose pas.

37. http://www.sd.be/website/be/fr/5000A/50C00C/50C10C/50C13C/10000P_120112_17_bedrijfswagens, page consultée le 1er février 2012.
38. http://tedsifflera3fois.com/2011/12/14/carnage-critique/, page consultée le 21 mars 2012.
39. http://www.mediapart.fr/journal/culture-idees/170910/pierre-rosanvallon-lechec-du-sarkozysme-la-panne-de-la-gauche?page_article=3, page consultée le 15 février 2012.

Certains linguistes ont réduit le problème de la position de l'adjectif à la sémantique liée à la position. Waugh (1977) et Bouchard (1998) ont chacun élaboré une théorie visant à rendre compte de la distribution des adjectifs épithètes en s'appuyant sur le rapport entretenu entre le nom et l'adjectif selon la position. Waugh (1977) adopte une approche structuraliste. Elle considère que chaque adjectif a un sens invariant et que c'est la position ainsi que le nom que l'adjectif modifie qui déterminent l'interprétation de l'adjectif en contexte. Dans le cas de la postposition, la combinaison du nom et de l'adjectif est conçue comme l'intersection de deux parties du discours différentes, alors que l'antéposition implique que le sens lexical du nom est présupposé.

Bouchard (1998) adopte un point de vue similaire, estimant, qu'en position prénominale, l'adjectif modifie des composants internes au nom, alors que postposé, l'adjectif modifie le nom pris comme un tout et lui assigne une propriété qui ne peut pas être attribuée à un sous-composant du nom [40]. Cet auteur appuie sa théorie notamment sur l'étude des adjectifs intensionnels *supposé*, *futur*, *parfait*…(cf. exemples (32), (33), (34)).

Ce type d'approche, visant à expliquer la position de l'adjectif par une différence de relation entre le nom et l'adjectif, oblige à postuler l'existence d'une différence sémantique systématique selon la position de l'adjectif. Or, une telle généralisation semble connaître des contre-exemples. D'abord, il existe des séquences Nom Adjectif qui ont un sens identique quelle que soit la position de l'adjectif, comme dans les syntagmes nominaux présentés en (88) (tirés de Abeillé & Godard, 1999, p. 28). De plus, dans certains cas, les différences sémantiques observées pour un couple Nom Adjectif ne sont pas valables pour n'importe quel couple Nom-Adjectif (89). Cela indique que ce n'est pas la position qui impose une interprétation à la combinaison Nom Adjectif, mais la combinaison elle-même. Enfin, comme nous l'avons montré par le biais d'exemples trouvés sur Google (exemples (38), (39), (40)), la sémantique particulière associée aux adjectifs intensionnels n'est pas spécifique à leur position, même si ces adjectifs présentent une préférence massive pour l'antéposition.

(88) a. *un charmant jeune homme*
b. *un jeune homme charmant*

(89) a. *un gros fumeur*
b. *un fumeur gros*
c. *un gros coiffeur*
d. *un coiffeur gros*

Dans l'exemple (88), il semble qu'il n'y ait pas de nuance sémantique permettant de distinguer une interprétation pour le syntagme avec adjectif antéposé ou post-

40. Plus exactement, l'analyse proposée par Bouchard se place dans le cadre du Programme Minimaliste et rend compte des différences sémantiques observées par une formalisation en termes de différences de relation syntaxique entre le nom et l'adjectif antéposé d'un côté, et le nom et l'adjectif postposé de l'autre. Les différences sémantiques sont alors conçues comme une conséquence de la différence dans la relation de modification syntaxique.

posé. L'exemple (89) fait apparaître que l'interprétation intensificatrice associée à l'antéposition de *gros* avec un nom comme *fumeur*, n'est pas valable en combinaison avec le nom *coiffeur* : *un gros coiffeur* renvoie à un individu qui est coiffeur et qui a la propriété d'être *gros*. De plus, comme le notent Abeillé & Godard (1999, p. 13) pour la séquence *un gros fumeur*, « *l'antéposition est compatible avec les deux interprétations* », à savoir qu'il s'agit soit d'un individu qui fume beaucoup, soit d'un individu qui est fumeur et gros.

Les différences sémantiques observées selon la position de l'adjectif ne sont donc pas systématiques et varient très fortement en fonction des items lexicaux combinés (adjectif et nom). Cependant, il existe bien des tendances que nous interprétons comme des contraintes préférentielles liées à l'adjectif lui-même et à la combinaison du nom et de l'adjectif.

En conclusion, nous considérons que la position de l'adjectif n'est pas associée à une sémantique spécifique. Le sens de l'adjectif, seul ou pris dans la séquence qu'il forme avec le nom, n'est qu'une contrainte préférentielle favorisant une position ou l'autre. Selon les adjectifs, les classes d'adjectifs et les combinaisons Nom Adjectif, ces préférences peuvent être très fortes ou quasiment inexistantes.

3.5.3. Stylistique

Reiner (1968) estime que le problème de la position de l'adjectif épithète est « *essentiellement d'ordre stylistique* » (Reiner, 1968, p. 4). Tous les adjectifs peuvent apparaître dans les deux positions et le choix d'une position se fait pour des raisons stylistiques. À travers l'inventaire des contraintes que nous avons effectué, nous avons montré que le problème ne peut pas se réduire à un aspect stylistique. Cependant, à partir du moment où il y a une certaine liberté du point de vue syntaxique, le locuteur a la possibilité de faire un choix stylistique. Nous estimons que les effets stylistiques se situent au niveau des choix individuels des locuteurs. Étant donné que nous nous intéressons aux tendances générales qui guident le choix de la position, les variations stylistiques individuelles n'entrent pas dans notre champ d'étude.

3.6. Effets de figements

Les adjectifs épithètes se rencontrent dans diverses expressions figées, comme en témoigne le livre de Gross (1996). Nous reproduisons, dans les exemples (90) à (92), des expressions figées extraites de cet ouvrage.

(90) noms

a. *canard* ***boiteux***
b. *bête* ***noire***
c. *tête* ***brûlée***
d. *poids* ***lourd***

(91) locutions adjectivales
   a. *dans le plus* ***simple*** *appareil*
   b. *d'âge* ***avancé***, *d'accès* ***difficile***, *d'un* ***seul*** *tenant*
   c. *en chute* ***libre***, *en* ***bonne*** *posture*
   d. *sur la corde* ***raide***
   e. *sous* ***haute*** *surveillance*

(92) locutions adverbiales
   a. *à bras* ***raccourcis***
   b. *de* ***longue*** *date*
   c. *de guerre* ***lasse***
   d. *tambour* ***battant***

De façon générale, la position de l'adjectif ne peut varier, dans la mesure où elle est fixée par l'expression dans laquelle apparaît l'adjectif en question. Deux remarques s'imposent en ce qui concerne ces effets de figements. Premièrement, la position de l'adjectif dans ces expressions semble respecter une position courante pour chacun des adjectifs. Ainsi, *long* et *bon* sont antéposés, tandis que *battant* ou *las* apparaissent en postposition. Deuxièmement, malgré leur caractère figé, certaines expressions offrent une souplesse quant à la position de l'adjectif. Parmi celles que nous avons énumérées précédemment, il semble que la postposition du SAdj soit possible dans l'exemple (91-a) comme cela est montré en (93). De même, l'exemple (94) atteste du fait que l'adjectif de (91-b) peut être antéposé, même s'il semble que la version précédemment citée est préférée.

(93) *dans l'appareil le plus* ***simple***

(94) *de* ***difficile*** *accès*[41]

Nous observons donc que la position de l'adjectif peut être dépendante d'effets de figement. Dans la plupart des cas, la position est imposée par l'expression utilisée, mais cette position est celle qui est généralement attendue, étant donné les préférences lexicales de chaque adjectif.

3.7. Aspects discursifs

Certains éléments touchant à l'organisation du discours peuvent avoir un effet sur le choix de la place de l'adjectif. Waugh (1977) mentionne que l'adjectif antéposé peut être utilisé dans des contextes anaphoriques : exemples (95) et (96), tirés de Waugh (1977, p. 132). Une fois établie la relation entre le référent du nom et la

41. Exemple de phrase contenant la séquence *difficile accès* : « *Ces textes ont été publiés dans des catalogues et des revues d'art de difficile accès, ou parfois seulement en traduction.* », http://www.fabula.org/actualites/h-cixous-peinetures-ecrits-sur-l-art_41766.php, page consultée le 14 septembre 2012.

propriété dénotée par l'adjectif, l'antéposition de l'adjectif est facilitée pour toute nouvelle occurrence du couple Nom Adjectif.

(95) *j'ai vu un éléphant* ***énorme****... Cet* ***énorme*** *éléphant buvait de l'eau*

(96) *C'est un livre bien écrit et important au point de vue littéraire, mais vénimeux et* ***méchant****... Eh bien je refuse absolument de faire lire ce* ***méchant*** *ouvrage à des élèves de quatorze ans*

Dans le contexte anaphorique, la séquence Adjectif - Nom a le même sens que la séquence Nom - Adjectif. Néanmoins, Waugh (1977) et Bouchard (1998), dont les théories reposent sur l'idée que chaque position est associée à un sens, estiment que la séquence anaphorique Adjectif - Nom confère une dimension déictique au référent désigné, ce que ne permet pas l'ordre inverse. Plus précisément, Waugh (1977) explique que « *la pré-position de l'adjectif suppose une recognition déictique du morphème lexical du substantif* » [42]. L'idée est que, lorsque la combinaison d'un nom et d'un adjectif est produite, le référent est construit. La séquence anaphorique Adjectif - Nom renvoie à ce référent connu et établi. La propriété dénotée par l'adjectif est alors présentée comme plus intimement liée au nom. Ces auteurs estiment que la séquence Nom - Adjectif en contexte anaphorique ne présente pas cette dimension déictique. Cela implique que la production de cette séquence ne fait pas appel à un référent connu et établi, mais présente le référent comme s'il était nouveau. Ainsi, dans l'exemple (97) (inspiré de Bouchard, 1998, p. 148), Bouchard (1998) et Waugh (1977) donnent une interprétation différente à (97-a) et à (97-b). Dans le premier cas, on ferait appel à un référent établi et connu, tandis que, dans le deuxième cas, le référent serait présenté comme s'il était nouveau.

(97) *Jean a une connaissance parfaite de ses capacités…*

a. *Cette* ***parfaite*** *connaissance de ses capacités lui permet de mieux vivre.*

b. *Cette connaissance* ***parfaite*** *de ses capacités lui permet de mieux vivre.*

Nous estimons qu'il n'y a pas de différence sémantique dans ce contexte et que le référent n'est pas présenté sous un angle différent en (97-a) et en (97-b). Cependant, nous admettons que le contexte anaphorique peut influencer la position de l'adjectif : la reprise anaphorique de la combinaison d'un nom et d'un adjectif favorise l'antéposition de l'adjectif. Notons que le contexte anaphorique impose l'utilisation du déterminant démonstratif (cf. section 3.4.5). Cela signifie que l'on s'attend à ce que le démonstratif favorise l'antéposition.

3.8. Quels adjectifs étudier ?

Le problème que nous avons décrit dans ce chapitre, ainsi que toutes les contraintes mises en jeu, impliquent l'existence d'une catégorie adjectif. Si l'appartenance à cette

42. « *pre-position of the adjective assumes the deictic recognition of the lexical morpheme of the substantive.* », (Waugh, 1977, p. 132-133)

catégorie des mots *petit*, *gros* et *joyeux* ne pose pas de problème, il existe plusieurs cas limites pour lesquels l'étiquette d'adjectif ne va pas de soi. De plus, le phénomène qui nous intéresse étant l'alternance de position des adjectifs, il est important de définir quels éléments sont pertinents de ce point de vue.

3.8.1. Les relationnels

La première question qui se pose lorsque l'on aborde la question de l'adjectif est de savoir si on parle seulement des adjectifs qualificatifs ou si l'on inclut les adjectifs relationnels. Les adjectifs relationnels se distinguent des adjectifs qualificatifs par leur comportement sémantique, syntaxique et morphologique. Cependant, la définition formelle de cette classe est problématique comme le montre par exemple Guyon (1993, chap. 2). Nous présentons les propriétés les plus importantes des adjectifs relationnels. Premièrement, au niveau morphologique, les adjectifs relationnels sont des dénominaux.

(98) a. *organisation / organisationnel*
b. *pétrole / pétrolier*

Deuxièmement, en termes syntaxiques, ils se distinguent des adjectifs qualificatifs par les fonctions qu'ils peuvent remplir. Alors que les qualificatifs peuvent être épithètes, apposés ou attributs du sujet ou de l'objet, les adjectifs relationnels ne peuvent occuper que la fonction épithète.

(99) *un problème organisationnel*

(100) ? *ce problème est organisationnel*

(101) ? *pétrolier, le puits a été creusé*

En plus des restrictions au niveau de la fonction, les relationnels ne peuvent pas être antéposés au nom(102), ni coordonnés (103), ni modifiés (104) par un adverbe.

(102) ? *un organisationnel problème*

(103) ? *un problème organisationnel et difficile*

(104) ? *un choc très pétrolier*

Troisièmement, au plan sémantique, les relationnels expriment une relation entre deux entités, et non la mise en relation d'une entité avec une propriété comme le font généralement les qualificatifs.

(105) *palais présidentiel* : exprime la relation entre un *palais* et un *président*

Les caractéristiques présentées ici ne permettent pas de définir une classe fermée. C'est pour cette raison que Goes (1999) définit le prototype d'un adjectif relationnel à partir de ses propriétés essentielles et classe ensuite les adjectifs selon le nombre de propriétés qu'ils partagent avec le prototype.

Au niveau morphologique, tous les dénominaux ne sont pas relationnels. Par exemple, l'adjectif *original*, dérivé de *origine*, peut avoir la fonction attribut (106-a) et être modifié par un adverbe (106-b) (exemples tirés de Guyon, 1993). De plus, certains relationnels ne sont pas dénominaux, comme *étudiant* dans le SN *le logement étudiant*.

(106) a. *ce voyage est original*
b. *un voyage très original*

Au niveau syntaxique, Bartning (1976) cite des contre-exemples en ce qui concerne la fonction attribut (107) et la modification adverbiale (108) avec les adverbes *strictement* et *purement*. Il existe également des exemples de relationnels modifiés par l'adverbe *très*, comme en (109).

(107) a. *les revendications en question sont syndicales*
b. *cette publication est mensuelle*

(108) a. *une revendication strictement féminine*
b. *des problèmes purement agricoles*

(109) *DDV, drapé dans une image d'homme audacieux au panache échevelé, en revient à une posture* ***très présidentielle*** *– incarnée sans interruption du Général de Gaulle à Jacques Chirac…* [43]

En ce qui concerne la coordination, Google fournit des exemples de relationnels coordonnés à un qualificatif, comme en (110). De même, l'antéposition d'un adjectif relationnel se rencontre (111), mais révèle une volonté de créer un effet stylistique de la part du locuteur/auteur.

(110) *Véritable problème* ***organisationnel et récurrent*** *de la politique sportive,* [44]

(111) *Seulement dans son subconscient, Sarko ne pouvait imaginer que cet homme, Fillon, qu'il qualifia par ce* ***présidentiel*** *mépris de "collaborateur", saurait mener aussi bien sa barque.* [45]

Nous donnons un exemple de phrase dans laquelle l'adjectif relationnel *présidentiel* est à la fois modifié, coordonné et antéposé.

(112) *l'intéressée […] a lancé un très présidentiel et inédit appel au rassemblement de toutes les forces de la gauche, jusqu'à l'extrême gauche.* [46]

43. http://leplus.nouvelobs.com/contribution/225115-villepin-candidat-a-la-presidentielle-il-a-tout-a-y-gagner.html, page consultée le 23 mars 2012.
44. http://www.comite-rhone-tennis.com/assets/_DOCS_/files/BASE%20DOCUMENTAIRE/COMMUNICATION/PLEINE%20LIGNE/25%20Pleine%20Ligne%20Juin%202011.pdf, page consultée le 23 mars 2012.
45. http://www.lepoint.fr/reactions/politique/commentaires-sur-remaniement-le-nouveau-bail-de-francois-fillon-a-matignon-14-11-2010-1261955_20, page consultée le 23 mars 2012.
46. http://actualite-generale.lalibre.be/_international/evenement-week-end.html, page consul-

Enfin, Guyon (1993, p. 65) donne l'exemple de l'adjectif *ultime*, qui partage des caractéristiques avec les relationnels et qui pourtant ne fait pas partie de cette classe. Cet adjectif ne peut pas avoir la fonction attribut, ni être modifié :

(113) a. ?*cette réponse est ultime*
b. ?*une réponse très ultime*

Dans la plupart des phrases que nous avons présentées comme contre-exemples aux caractéristiques syntaxiques des relationnels, on peut estimer que les adjectifs ont perdu leur dimension relationnelle, c'est-à-dire qu'ils n'expriment plus une relation. Dans ce cas, on ne peut pas définir une classe lexicale de relationnels, mais on doit parler d'une classe d'emplois relationnels qui se caractérisent notamment par les propriétés syntaxiques et sémantiques définies plus haut.

Les exemples fournis précédemment montrent que les adjectifs dits relationnels peuvent, dans des conditions spécifiques, se comporter comme des qualificatifs au niveau syntaxique. De plus, les adjectifs qualificatifs n'ont pas un comportement syntaxique homogène en ce qui concerne la fonction syntaxique et la modification adverbiale (exemple de *ultime*). Nous considérons que l'étude des adjectifs dits relationnels est pertinente dans la mesure où ils peuvent être antéposés. Tout comme les adjectifs de nationalité, ces adjectifs forment une sous-classe d'adjectifs qui a une très forte préférence pour la postposition et qui, dans des conditions d'emploi particulières, peut se rencontrer en antéposition.

3.8.2. Les ordinaux

Les ordinaux forment un paradigme spécifique d'adjectifs apparaissant très massivement en antéposition. Deux cas particuliers se distinguent : *premier* et *second*. L'adjectif *second* peut être postposé dans des expressions telles que *langue seconde*, *propriété seconde*, *réalité seconde*, *personnage second*. Son sens se rapproche alors de celui de l'adjectif *secondaire*. De même, l'adjectif *premier* peut apparaître postposé au nom. Il porte alors le sens de *nécessaire/essentiel* (114). Ce sens peut également se rencontrer lorsque l'adjectif est antéposé (115).

(114) a. *devoirs premiers*
b. *condition première*
c. *nécessité première*

(115) *(de) première nécessité*

Les autres adjectifs ordinaux ne peuvent apparaître en postposition que dans le cas spécifique où ils sont associés à un nom désignant une partie d'ouvrage (*chapitre*, *livre*…). Il s'agit d'un usage littéraire figé.

En nous appuyant sur ces données, nous estimons que les adjectifs *premier* et *second* sont pertinents pour la question de la position de l'adjectif. En revanche, les autres

tée le 23 mars 2012.

ordinaux ne présentent pas de possibilité de mobilité "productive", à savoir qu'une séquence autre que *livre deuxième*, *chapitre deuxième* ou *tome deuxième* ne peut pas être produite. Nous les écartons donc de notre objet d'étude.

3.8.3. Les indéfinis

Les adjectifs indéfinis constituent une catégorie intermédiaire entre les déterminants indéfinis et les adjectifs qualificatifs. Les éléments souvent regroupés sous cette étiquette n'ont pas un comportement homogène. Premièrement, nous réduisons la catégorie des indéfinis aux adjectifs apparaissant avec un déterminant, ce qui permet notamment d'écarter *chaque*, *nul* et *aucun*. De plus, nous limitons cette classe aux adjectifs pouvant se présenter en postposition, comme dans les exemples (116) à (121). Les indéfinis sont alors au nombre de 6 : *différent*, *autre*, *certain*, *quelconque*, *divers*, *tel*. En ce qui concerne les adjectifs *tel* (121) et *certain* (118), les phrases rencontrées peuvent apparaître à la limite de l'acceptable. Cependant, ce type de phrases n'est pas rare, ce qui tend à montrer que ces adjectifs ont une mobilité certaine.

(116) *L'approche mimétique peut-elle nous permettre de considérer ensemble et, dans leur mouvement d'engendrement, les phases* ***différentes*** *des processus de la violence comme de leur possible issue.* [47]

(117) *Pour toute situation* ***autre****, les familles déposent obligatoirement une demande de dérogation.* [48]

(118) *Les enfants ont pu découvrir une approche* ***certaine*** *de ce qu'est la vie en collectivité.* [49]

(119) *Juste une bête adaptation avec une histoire* ***quelconque*** *pour jeunes japonais dépravés en manque d´animé.* [50]

(120) *Depuis que j'ai commencé à écrire ce blog, je ne compte plus les sollicitations* ***diverses*** *de journalistes m'enjoignant de partager mes "bons plans", "mes trucs et astuces" de mère de famille nombreuse...* [51]

(121) *il vous explique pas à pas les démarches pour constituer un dossier de surendettement, de la préparation de votre dossier, à son passage en commission, et vous livre des conseils judicieux pour éviter de vous retrouver dans une situation* ***telle****.* [52]

47. http://home.nordnet.fr/~jpkornobis/Textes/mlme.htm, page consultée le 23 mars 2012.
48. http://www.ac-grenoble.fr/ia26/spip/IMG/pdf_annexe_4_2011_List_DPT_limitrophe.pdf, page consultée le 23 mars 2012.
49. http://centredeloisirslivarot.over-blog.com/article-le-centre-a-referme-ses-portes-vendredi-que-de-souvenirs-82615421.html, page consultée le 23 mars 2012.
50. http://www.jeuxvideo.com/forums/1-18-9031155-533-0-1-0-0.htm, page consultée le 23 mars 2012.
51. http://www.lacavernedelala.com/fr/blog/52-les-bons-plans-de-la-famille-dejantee, page consultée le 23 mars 2012.
52. http://www.infinance.fr/web/credit/credit-a-la-consommation/page-web-comment-faire-un-dossier-de-surendettement-2790.htm, page consultée le 23 mars 2012.

Enfin, nous abordons le cas de l'indéfini *même*. Cet adjectif peut apparaître en antéposition (122-a) et en postposition (122-b). Cependant, alors que le sens de l'adjectif antéposé peut se paraphraser par *identique*, l'adjectif postposé exprime l'ipséité. Il s'agit donc deux lemmes distincts.

(122) a. *la même période*
b. *la période même*

Il apparaît que pour ces deux homonymes, il n'est pas possible de trouver le sens *identique* en postposition ou bien le sens de l'ipséité en antéposition, et ce quel que soit le nom utilisé. Cela indique que c'est bien la position, et non l'item nominal modifié, qui définit le sens pour ces deux homonymes. Étant donné que l'alternance de position n'est pas possible, nous écartons ces deux homonymes de notre étude.

Après avoir dressé, d'après la bibliographie sur le sujet, le bilan des contraintes intervenant dans le choix de la position de l'adjectif épithète, nous récapitulons les contraintes préférentielles sous forme de liste :

Phonologie

défectivité pour la FLMS	favorise la postposition
hiatus dans la séquence Nom - Adjectif	favorise l'antéposition

Lexique

longueur relative	favorise l'ordre *court avant long*
longueur absolue	favorise l'antéposition des monosyllabiques favorise la postposition des polysyllabiques
fréquence	favorise l'antéposition des lemmes très fréquents favorise la postposition des lemmes moins fréquents
dénominaux et déverbaux	favorise la postposition
préfixe privatif *in-*	favorise l'antéposition
catégories "objectives"	favorise la postposition
catégorie des subsectifs	favorise l'antéposition
catégorie des évaluatifs	pas de préférence apparente, particulièrement mobile

Syntaxe

modifieur pré-adjectival	favorise la postposition (effet de longueur) augmente la mobilité d'adjectifs lexicalement contraints
coordination	favorise la postposition (effet de longueur) augmente la mobilité d'adjectifs lexicalement contraints
présence d'autres dépendants du nom	favorise l'antéposition
déterminant indéfini	favorise la postposition
déterminant défini	favorise l'antéposition
SN sujet	favorise l'antéposition
SN attribut et déterminant défini	favorise l'antéposition

Sémantique

adjectifs homonymes	une position est largement favorisée par un sens
couple nom - adjectif	conditionne la position de l'adjectif

Discours

reprise anaphorique du SN	favorise l'antéposition

L'ensemble de ces contraintes peut donc être envisagé comme une série de contraintes préférentielles. Dans le chapitre suivant, nous étudierons le comportement d'une partie importante de ces contraintes sur des données de corpus et nous tâcherons de formaliser leur comportement à l'aide des outils décrits dans le chapitre 2.

Chapitre 4

Analyse de données de corpus

Sommaire

L'objectif de ce chapitre est de présenter un travail mené sur corpus dans le but d'analyser le phénomène de la position de l'adjectif épithète. Dans un premier temps, nous décrirons les méthodes utilisées pour extraire les données pertinentes. Dans un deuxième temps, nous présenterons les variables qui ont été annotées sur ce corpus et qui permettront l'étude des contraintes. Dans la dernière partie, nous exposerons la modélisation du phénomène, faite à partir des contraintes étudiées. Nous présenterons différents modèles statistiques construits sur différents faisceaux de contraintes. En nous appuyant sur leur comparaison, nous proposerons une estimation de l'importance relative des divers aspects linguistiques que nous avons décrits dans le chapitre précédent. Enfin, nous exposerons le bilan de nos modèles sur corpus, ainsi qu'une étude corrélationnelle, s'appuyant sur un questionnaire d'élicitation de préférences, que nous avons mise en place dans le but de montrer que les observations faites en corpus recoupent les observations du comportement langagier des locuteurs.

4.1. Extraction des données

Les données ont été extraites de la sous-partie annotée en fonctions du *French Treebank* (FTB) [1], corpus arboré d'articles du journal *Le Monde*. À partir de l'annotation morphologique et syntaxique de ce corpus, nous avons extrait automatiquement les mots catégorisés 'adjectif' modifiant une tête lexicale catégorisée 'nom'. Notons que les éléments annotés comme mots composés ont été traités comme des formes uniques non structurées. Étant donné que la structure morphologique associée aux mots composés est plus pauvre que celle utilisée pour les constituants syntaxiques, l'utilisation de cette structure aurait pu induire un certain nombre d'erreurs. Les mots composés tels que *petit-fils*, *grand père* ou *bon marché*, ont donc été écartés. Les données ainsi extraites comptent 17 762 occurrences d'adjectifs épithètes.

4.1.1. Nettoyage de la table

Dans le FTB, les numéraux cardinaux sont annotés comme adjectifs, lorsqu'ils apparaissent avec un déterminant ou dans une date [2]. Ces éléments ne sont pas pertinents pour notre étude. Nous les avons éliminés, ainsi que les ordinaux, excepté *premier* et *second* pour les raisons décrites dans le chapitre précédent. De plus, étant donné que les composés ont été considérés comme des formes uniques, certains éléments annotés comme adjectifs sont en fait des locutions construites à partir d'un SP ou d'un SN. Par exemple, le SP *en fin de droits* de l'exemple (1) et la séquence *aller et retour* en (2) sont catégorisés comme des adjectifs [3].

(1) *des chômeurs indemnisés en_fin_de_droits*

1. Pour plus de détails sur ce corpus arboré, voir la section 2.1.3 du chapitre 2.
2. Exemple d'annotation de date dans le FTB : *(NP (DET Le) (NC jeudi) (ADJ 13) (NC mai) (NC 1993))*.
3. Nous signalons les éléments formant un mot composé en les reliant à l'aide du signe '_'.

(2) *les billets aller_et_retour*

De même, les noms ayant un emploi épithétique, ou substantifs épithètes selon les termes de Noailly (1990), sont parfois étiquetés adjectifs, comme dans les exemples présentés en (3).

(3) a. *(NP (DET leur) (NC voyage) (AP (ADJ **éclair**))*
b. *(NP (DET la) (NC recette) (AP (ADJ **miracle**))…)*
c. *(NP (DET un) (NC rôle) (AP (ADJ **clé**)))*

Dans la mesure où notre travail est axé sur la catégorie de l'adjectif, les substantifs épithètes et les mots composés étiquetés adjectif ont été écartés de la table de données. De plus, nous avons éliminé les emprunts tels que *high-tech*, *junk*, *offshore* ou encore *ejidales*.

Les données extraites automatiquement contiennent également des sigles, tels que *PIB*, *OPA*, *OPE*, ainsi que leur équivalent sous forme non abrégée *produit intérieur brut*, *offre publique d'achat*, *offre publique d'échange*. Dans ce cas, les mots composant l'expression sont analysés comme des adjectifs et des noms. Afin d'harmoniser au maximum les donnés étudiées, nous avons fait le choix d'exclure ces occurrences de notre table de données. Dans le même ordre d'idées, les adjectifs apparaissant dans des expressions renvoyant à des entités nommées, telles que des noms d'entreprises (4), d'organisations (5) ou encore de journaux (6), ont été éliminés.

(4) a. *Les ciments français*
b. *Groupement Foncier Français*

(5) a. *Fonds Monétaire International*
b. *Conseil d'Etats indépendants*

(6) a. *Le Nouvel Observateur*
b. *Bonne Soirée*

Enfin, nous avons nettoyé la table de données des occurrences qui ne correspondaient pas à notre objet de recherche (7-a) ou qui présentaient des problèmes d'annotation (7-b). À chaque fois que nous avons rencontré ce type d'éléments, nous l'avons supprimé manuellement.

(7) a. *(NP (ADJ triple) (NC A))*
b. *(NP (DET ses) (NC activités) (NC grands_travaux) (AP (ADJ international)))*

4.1.2. Dépendants postadjectivaux et homonymes

La présence d'un dépendant postadjectival impose la postposition du SAdj. Le placement des adjectifs accompagnés d'un dépendant répond donc à une contrainte dure. Ces données ne sont pas pertinentes pour l'étude de préférences que nous nous proposons de mener. Nous avons éliminé l'ensemble des SAdj dans lesquelles l'adjectif

était suivi d'un dépendant, ce qui correspondait à environ 500 occurrences.

Dans le chapitre précédent, nous avons postulé l'existence d'adjectifs homonymes pour lesquels il faut distinguer deux entrées lexicales. La lemmatisation du FTB ne fait pas la distinction entre les différents homonymes d'une même forme. Nous avons donc différencié manuellement les paires suivantes :

- *ancien1* et *ancien2*,
- *cher1* et *cher2*,
- *commun1* et *commun2*,
- *pauvre1* et *pauvre2*,
- *propre1* et *propre2*,
- *pur1* et *pur2*,
- *sacré1* et *sacré2*,
- *seul1* et *seul2*,
- *simple1* et *simple2*

4.1.3. La table de données

Les données recueillies et nettoyées sont stockées sous la forme d'une table de données. Chaque ligne de la table correspond à une occurrence adjectivale. Elle contient les formes adjectivales et nominales, les lemmes correspondants et la position attestée de l'adjectif par rapport au nom. La position de l'adjectif correspond à la variable à prédire. Cette variable s'appelle `position`. Elle prend la valeur 0 quand l'adjectif est postposé au nom, et la valeur 1 quand il est antéposé. Chaque contrainte préférentielle, aussi appelée variable prédictrice, est représentée sous la forme d'une colonne. Un exemple de table de données est présenté dans la table 4.1. La variable `construit` indique si l'adjectif est morphologiquement construit (1) ou non (0). La variable `longAbs` donne la longueur de l'adjectif en nombre de syllabes.

adj	nom	lemme_A	lemme_N	position	construit	longAbs
espagnole	filiale	espagnole	filiale	0	1	3
difficile	évolution	difficile	évolution	0	0	3
première	caisse	premier	caisse	1	0	2
nouvelle	vigueur	nouveau	vigueur	1	0	2

Table 4.1.: Un exemple de table de données

Notre table de données contient 13933 occurrences d'adjectifs qui se répartissent en 3809 occurrences antéposées et 10124 postposées, soit 72.7% de postposition. Ce chiffre est supérieur aux taux de postposition rencontrés chez Wilmet (1981) et chez Forsgren (1978) (respectivement 67.2% et 66.4%). Les données de Wilmet ne relèvent pas du genre journalistique, mais du genre littéraire. On peut supposer que la proportion plus élevée d'antéposition est, en partie, due au genre du corpus, dans la mesure où le genre littéraire semble privilégier la position pré-nominale par rapport aux autres genres de discours. Le corpus de Forsgren rassemble des données extraites de journaux, tout comme notre table de données. Cependant, Forsgren a éliminé de ses données les adjectifs dits relationnels, de nationalité et de couleur. Étant donné l'importance des adjectifs de nationalité et de type relationnel dans nos

données, on peut supposer que c'est l'absence de ces adjectifs qui rend la proportion de postposition plus faible chez Forsgren.

Les 13933 occurrences de la table de données représentent 1750 lemmes adjectivaux. Parmi ces lemmes, 1487 n'apparaissent qu'en postposition, 92 qu'en antéposition et 171 dans les deux positions. En d'autres termes, près de 90% des lemmes n'apparaissent que dans une seule position. Cependant, le nombre réduit de lemmes présentant une alternance dans nos données représente un nombre élevé d'occurrences. Ces données sont résumées dans la table 4.2. Le ratio entre le nombre d'occurrences et le nombre de lemmes dans les trois groupes montre bien que les lemmes se présentant dans une position unique sont peu fréquents en comparaison aux lemmes apparaissant dans les deux positions.

	Antéposés uniquement		Postposés uniquement		2 positions		Totaux	
Lemmes	92	5.2%	1487	85.0%	171	9.8%	1750	100%
Occurrences	462	3.3%	8477	60.8%	4994	35.8%	13933	100%
Ratio occurrences / lemmes	5.0		5.7		29.2		8.0	

TABLE 4.2.: Lemmes et occurrences en antéposition, en postposition et dans les deux positions

Pour terminer cet aperçu général des données, notons que les adjectifs apparaissant dans les deux positions sont à 67.0% antéposés. Les adjectifs alternant effectivement dans nos données sont donc fréquents et préfèrent l'antéposition. Cela est cohérent avec l'idée que plus un lemme est fréquent, plus il a tendance à être antéposé. Nous reviendrons sur ce point dans la section 4.2.2.

Dans la section suivante, nous présentons les contraintes étudiées. Nous exposons la façon dont ces variables ont été obtenues et codées, ainsi que quelques éléments descriptifs permettant d'estimer leur impact sur la variable à prédire, c'est-à-dire la position de l'adjectif.

4.2. Les contraintes préférentielles étudiées

Nous présentons d'abord les contraintes concernant l'item adjectival (longueur, fréquence, morphologie et classe lexicale), puis celles relatives à la syntaxe (modification, coordination, autres éléments présents dans le SN, déterminant, fonction du SN). Nous aborderons ensuite les contraintes se rapportant à la combinaison du nom et de l'adjectif. Enfin, nous ferons le point en ce qui concerne la liaison et le hiatus.

Les données que nous présentons en termes de nombre d'occurrences et de proportions ont pour but de donner un aperçu relativement précis du contenu de notre table de données. Cependant, il est difficile d'interpréter ces proportions, dans la mesure où le problème que nous traitons est multifactoriel et qu'il faut donc prendre en compte

l'ensemble des variables pour avoir une meilleure image de l'effet de chaque variable. C'est ce que nous permettra la modélisation statistique que nous détaillerons dans la section 4.3.

4.2.1. Longueur

Les contraintes relatives à la longueur sont au nombre de trois :

1. longueur absolue : plus l'adjectif est court, plus il a tendance à être antéposé ;
2. longueur relative : si l'adjectif est plus court que le nom, il a tendance à être antéposé, et dans le cas inverse, il a tendance à être postposé ;
3. longueur du SAdj : plus le SAdj est long, plus il a tendance à être postposé.

Afin de capter ces trois contraintes, nous avons estimé la longueur en syllabes de l'adjectif, du nom et du SAdj. Pour cela, nous avons utilisé le logiciel industriel de synthèse vocale eLite [4]. La syllabation ainsi effectuée tient compte des effets de liaison. À partir de ces comptes de syllabes, nous avons créé trois variables captant les trois contraintes à étudier :

`longAbs` : nombre de syllabes de l'adjectif ;

`longRel` : nombre de syllabes de l'adjectif moins nombre de syllabes du nom ;

`longSAdj` : nombre de syllabes du SAdj.

La longueur codée par ces variables correspond à la longueur moyenne du mot dans l'ensemble du corpus syllabé. Par exemple, le logiciel de syllabation assigne trois syllabes au mot *terrible* dans un contexte où il n'y a pas de liaison (*terrible pression*) et deux syllabes dans un contexte de liaison (*terrible enchaînement*). La longueur de *terrible* dans la table de données correspond à la moyenne de ces longueurs. Cela signifie que, pour *terrible*, `longAbs` = 2.5. Les variables de longueur ne prennent donc pas seulement des entiers comme valeurs.

Les statistiques descriptives relatives aux trois variables, `longAbs`, `longSAdj` et `longRel`, sont présentées dans le tableau 4.3. On observe que les adjectifs de la table de données comportent entre 1 et 9 syllabes et que leur longueur moyenne est de 2.63 syllabes. Les SAdj ont une longueur maximale de 40.6 syllabes et une moyenne proche de 3. Enfin, la moyenne (0.04) de la longueur relative `longRel` indique que le nom et l'adjectif tendent à avoir la même longueur.

Ces trois variables semblent avoir une influence sur la position de l'adjectif. Cependant, elles sont très corrélées entre elles, dans la mesure où `longRel` est calculée à partir de `longAbs` et que `longSAdj` est égale à `longAbs` dans tous les cas où l'adjectif est le seul constituant du SAdj. Il faut déterminer si chaque variable aide réellement à décrire le comportement de la variable `position`, indépendamment des deux autres. De cette façon, on pourra avoir une idée plus précise de l'importance de l'effet de chacune des trois variables. Dans ce but, nous présentons et comparons les variables `longAbs` et `longRel` dans un premier temps, puis nous mettrons en parallèle

4. Ce logiciel est vendu par Multitel asbl à Mons (Belgique). Nous remercions Richard Beaufort, Sophie Roeckhoudt et Benoît Crabbé pour leur aide à la réalisation de la syllabation.

	Min.	Médiane	Max.	Moyenne	Ecart-type
`longAbs`	1	2.5	9	2.63	1.03
`longSAdj`	1	3	40.6	2.98	1.63
`longRel`	−6	0	7	0.04	1.52

TABLE 4.3.: Statistiques descriptives concernant les variables `longAbs`, `longRel` et `longSAdj`

les variables `longAbs` et `longSAdj` dans un deuxième temps. Nous terminerons cette section relative à la longueur en comparant les mesures en nombre de syllabes et en nombre de caractères.

`longAbs` et `longRel`

La corrélation entre la longueur absolue et la position de l'adjectif est claire. Les adjectifs de plus de deux syllabes sont postposés dans l'immense majorité des cas (92.6%), alors que les adjectifs de moins de deux syllabes sont très largement antéposés, dans 73.3% des cas. La valeur seuil est de deux syllabes : les adjectifs bisyllabiques sont plus souvent postposés, mais on les rencontre en antéposition dans 36.1% des cas. Ces chiffres sont présentés dans la table 4.4.

	Antéposés		Postposés		Totaux		Nombre de lemmes
`longAbs` < 2	1567	73.3%	571	26.7%	2138	100%	123
`longAbs` $= 2$	1722	36.1%	3045	63.9%	4775	100%	429
`longAbs` > 2	520	7.4%	6509	92.6%	7030	100%	1216

TABLE 4.4.: La variable `longAbs` en fonction de `position`

Il existe aussi une relation entre la variable `longRel` et la position de l'adjectif, mais, comme dans les données de Wilmet (1981) (cf. partie 3.3.1, chapitre 3), la tendance est moins forte que pour la longueur absolue. Dans la table 4.5, on observe que lorsque l'adjectif et le nom sont de même longueur (`longRel` $= 0$), les proportions suivent les proportions de la table de données entière. Lorsque l'adjectif est plus court (`longRel` < 0), le taux d'antéposition augmente pour atteindre 43.6%. Inversement, lorsque l'adjectif est plus long que le nom (`longRel` > 0), c'est le taux de postposition qui augmente jusqu'à 86.1%.

	Antéposés		Postposés		Totaux	
`longRel` < 0	2250	43.6%	2916	56.4%	5166	100%
`longRel` $= 0$	764	25.2%	2271	74.8%	3035	100%
`longRel` > 0	795	13.9%	4937	86.1%	5732	100%

TABLE 4.5.: La variable `longRel` en fonction de `position`

Si la longueur relative permet de mieux décrire la position de l'adjectif, c'est que la longueur du nom apporte une information supplémentaire à celle de la longueur de l'adjectif. Une rapide observation du comportement de la longueur du nom en nombre de syllabes par rapport à la variable `position`, montre que ce n'est pas le cas. Que le nom soit court (moins de 2 syllabes), moyen (2 syllabes) ou long (plus de 2 syllabes), la proportion de postposition reste autour de 72%. Étant donné que la longueur du nom n'est pas pertinente, on peut supposer que l'effet de la variable `longRel` n'est que la conséquence de l'effet de la variable `longAbs` : quand l'adjectif fait moins de deux syllabes, il est plus souvent plus court que le nom, et inversement, quand l'adjectif a plus de deux syllabes, il y a plus de chances qu'il soit plus grand que le nom.

En conclusion, la longueur du nom n'intervient pas dans le choix de la position. La différence de longueur entre l'adjectif et le nom n'est significative que grâce à l'effet de la longueur de l'adjectif. La longueur relative n'est donc pas pertinente. Seule la longueur absolue est une contrainte appropriée à la description de la position de l'adjectif.

`longAbs` et `longSAdj`

Le comportement de la variable `longSAdj` est très semblable à celui de `longAbs`, comme le montrent les données répertoriées dans la table 4.6.

	Antéposés		Postposés		Totaux	
`longSAdj` < 2	1362	77.8%	389	22.2%	1751	100%
`longSAdj` $= 2$	1709	38.2%	2767	61.8%	4476	100%
`longSAdj` > 2	738	9.6%	6968	90.4%	7706	100%

TABLE 4.6.: La variable `longSAdj` en fonction de `position`

Les variables `longAbs` et `longSAdj` sont très fortement corrélées. Lorsque l'adjectif apparaît seul dans le SADJ, c'est-à-dire dans 87.6% des cas, les valeurs de ces variables sont identiques. Les deux variables sont donc très redondantes.

Pour conserver l'information apportée par la variable `longSAdj` tout en neutralisant la corrélation, nous créons une nouvelle variable `ratioLong` qui prend pour valeur le rapport `longSAdj/longAbs`. Cette variable permet de savoir si le SADJ est plus long que l'adjectif seul. Si `ratioLong` $= 1$, l'adjectif est le seul composant du SADJ. Au-delà de 1, plus la valeur de `ratioLong` est élevée plus le SADJ est long par rapport à l'adjectif seul. [5]

5. Le tableau suivant contient les statistiques descriptives relatives à la variable `ratioLong`.

	Min.	Médiane	Max.	Moyenne	Ecart-type
`ratioLong`	1	1	20.31	1.15	0.6

La table 4.7 donne le comportement de la variable `position` en fonction de la variable `ratioLong`. Lorsque le SAdj est plus de deux fois plus grand que l'adjectif, on observe la postposition à plus de 90%.

	Antéposés		Postposés		Totaux	
`ratioLong` = 1	3406	27.9%	8795	72.1%	12201	100%
1 > `ratioLong` >= 2	352	32.1%	745	67.9%	1097	100%
`ratioLong` > 2	584	8.0%	51	92.0%	635	100%

Table 4.7.: La variable `ratioLong` en fonction de `position`

Les variables `longAbs` et `ratioLong` permettent respectivement de capter la longueur de l'adjectif et la différence de longueur entre l'adjectif et le SAdj. L'avantage d'utiliser ces deux variables est qu'elles ne sont quasiment pas corrélées. Afin d'évaluer la corrélation des variables, nous utilisons le ρ_s de Spearman, qui est une mesure de corrélation sur des données rangées et qui ne présuppose donc pas la linéarité de la relation. Pour `longAbs` et `ratioLong`, $\rho = -0.06$ (p= 2.4×10^{-11}). Cela indique que la corrélation est quasi inexistante. Ces deux variables expliquent donc chacune, de façon indépendante, une part de la variation de la variable `position`.

Longueur en syllabes et en nombre de caractères

En plus de la variable `longAbs` qui permet de connaître la longueur de l'adjectif en nombre de syllabes, nous avons créé une variable, `longCaract`, qui donne la longueur de l'adjectif en nombre de caractères. On observe que les mesures en syllabes et en caractères sont fortement corrélées : $\rho = 0.84$ (p< 2.2×10^{-16}). Les effets de ces deux mesures de longueur sur la variable `position` sont relativement identiques. Afin de pouvoir comparer les deux mesures, nous les avons standardisées en utilisant le score z, ce qui revient à ce que les deux variables aient une moyenne de 0 et un écart-type de 1 (cf. section 2.2.1.2.3 du chapitre 2) [6]. C'est à partir de ces variables standardisées que nous avons construit le graphique de la figure 4.1, qui représente la longueur de l'adjectif en fonction de la proportion d'antéposition. La relation entre longueur et position est rendue explicite par les courbes de régression : celle en rouge est relative au nombre de caractères, celle en bleu au nombre de syllabes [7].

6. La variable `longAbs` a également été arrondie à 0.5 près dans le but de réduire le nombre de points sur le graphique.

7. Les courbes sont les représentations graphiques des modèles de régression logistique suivants :

```
             Estimation  Erreur-type  valeur z  Pr(>|z|)
(Intercept)  -1.40711    0.02667      -52.76    <2e-16
longAbs      -1.50420    0.03205      -46.93    <2e-16

             Estimation  Erreur-type  valeur z  Pr(>|z|)
(Intercept)  -1.24809    0.02357      -52.94    <2e-16
longCaract   -1.18157    0.02776      -42.57    <2e-16
```

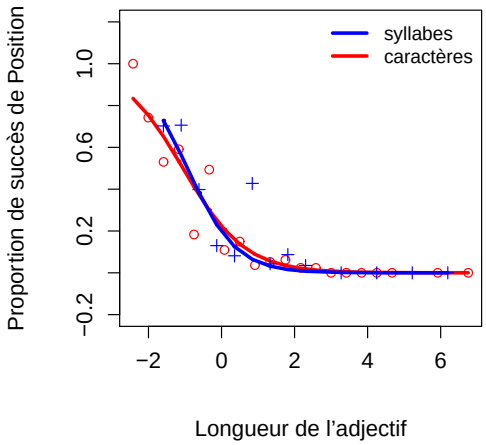

FIGURE 4.1.: Longueur de l'adjectif en syllabes (bleu) et en nombre de caractères (rouge) en fonction de `position` avec les courbes logistiques résumant le mieux les données.

Ce graphique montre que les mesures de longueur en syllabes et en caractères semblent relativement équivalentes par rapport au problème de la position de l'adjectif épithète. Cela signifie que le nombre de caractères, qui est une mesure facile à obtenir, semble suffisant pour rendre compte de l'effet de la longueur sur la position de l'adjectif. Dans la suite de notre travail, nous utiliserons la mesure syllabique.

4.2.2. Fréquence

Wilmet (1981) considère que la fréquence est un facteur plus pertinent que la longueur. Dans cette partie, nous décrirons le comportement de la variable position en fonction d'une variable mesurant la fréquence, puis nous ferons un point sur la relation entre longueur et fréquence dans nos données.

La fréquence d'un lemme peut être envisagée comme un estimateur de la probabilité d'emploi de ce lemme : plus la fréquence est haute, plus la probabilité d'utiliser le mot est élevée. Nous avons utilisé deux ressources pour estimer au mieux la fréquence : le corpus de l'Est-Républicain (ER)[8] et la base de données *Lexique 3.72* (New, 2006; New *et al.*, 2001)[9]. Le corpus ER est un corpus d'articles de journaux, qui contient 148 millions de mots, ce qui laisse supposer que c'est un bon estimateur de fréquence pour le genre journalistique. Afin de ne pas limiter la fréquence au

8. Pour plus de détails sur ce corpus et son annotation en parties du discours, voir la section 2.1.3 du chapitre 2.

9. Ce lexique est disponible sur le site : http://www.lexique.org/.

genre journalistique, nous avons ajouté les fréquences disponibles dans *Lexique 3.72*. Ce lexique contient 142 694 formes et 46 947 lemmes, dont 26 806 formes adjectivales et 11 580 lemmes adjectivaux. Les données relatives à la fréquence sont issues, d'une part, de 218 romans publiés entre 1950 et 2000, qui représentent 14.7 millions d'items (corpus littéraire) et, d'autre part, de sous-titres de 9 474 films ou saisons de séries représentant en tout 50 millions de mots (corpus de sous-titres). La fréquence que nous utilisons est donc l'addition de la fréquence dans ER, de la fréquence dans le corpus de sous-titres et de la fréquence littéraire.

La mesure exacte que nous utilisons pour le corpus journalistique correspond à la fréquence d'apparition de chaque adjectif en position épithète dans le corpus ER. Ce corpus n'étant pas annoté en constituant, nous avons considéré qu'un adjectif épithète est un mot étiqueté Adjectif adjacent à un mot étiqueté Nom [10]. La fréquence retenue pour ce corpus est le nombre d'occurrences du lemme en position épithète divisé par 148, pour avoir une fréquence exprimée en nombre d'occurrences par million de mots (format utilisé dans *Lexique 3.72*). Les mesures exactes pour les corpus littéraire et sous-titres sont les valeurs pour les lemmes ayant la catégorie adjectif. La variable `freq` correspond donc à la fréquence par million sur les trois corpus [11]. On observe 60 adjectifs avec une fréquence nulle et la fréquence la plus élevée est de 1029, pour l'adjectif *petit*. Les statistiques descriptives concernant la fréquence sont présentées dans la table 4.8.

	Min.	Médiane	Max.	Moyenne	Ecart-type
`freq`	0	52.18	1029	183.3	289.0

TABLE 4.8.: Statistiques descriptives concernant la variable `freq`

Afin d'illustrer l'effet de la variable `freq`, nous avons fait quatre groupes de taille à peu près équivalente selon la fréquence, comme cela est montré dans la table 4.9. De façon logique, les lemmes les plus fréquents sont les moins nombreux. On observe également que les adjectifs les moins fréquents favorisent très majoritairement la postposition, tandis que les plus fréquents préfèrent l'antéposition à 64.2%. Entre les deux, les proportions de postposition ont tendance à augmenter à mesure que les adjectifs sont moins fréquents.

Le graphique de la figure 4.2 représente la fréquence de l'adjectif en fonction de la proportion d'antéposition. Afin d'éviter que le nuage de points soit trop dispersé, nous avons arrondi les valeur de `freq` à 1 près. La courbe obtenue est la courbe de

10. Le repérage de la fonction épithète dans ER se fait au moyen d'une approximation qui implique des erreurs, notamment dans le cas d'adjectif attribut de l'objet :

(i) Luc boit son [rhum]$_N$ [chaud]$_{Adj}$

11. Plus précisément, nous avons additionné, pour chaque lemme, les fréquences brutes provenant des trois corpus. Nous avons ensuite divisé cette fréquence par 214.7, nombre de millions de mots des trois corpus réunis.

	Antéposés		Postposés		Totaux		Nombre de lemmes
`freq` >= 194.2	2340	66.7%	1169	33.3%	3509	100%	28
194.2 > `freq` >= 52.18	910	26.5%	2529	73.5%	3439	100%	112
52.18 > `freq` >=10.7	410	11.6%	3115	88.4%	3525	100%	329
10.7 > `freq`	149	4.3%	3311	95.7%	3460	100%	1288

TABLE 4.9.: La variable `freq` en fonction de `position`

régression la mieux ajustée aux données[12]. La relation entre fréquence et position apparaît clairement : les adjectifs ayant une fréquence limitée sont majoritairement postposés (proportion d'antéposition entre 0 et 0.5), tandis que les adjectifs plus fréquents préfèrent l'antéposition (proportion d'antéposition entre 0.5 et 1).

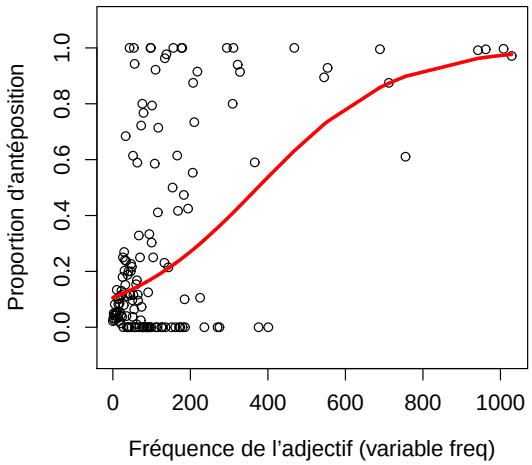

FIGURE 4.2.: `freq` en fonction de `position` avec la courbe logistique résumant le mieux les données.

Depuis les travaux de Zipf (1932), on sait qu'il existe une relation inverse entre la fréquence et la longueur des mots : plus un mot est fréquent plus il a tendance à être court. Cela s'explique en partie par la simplification de l'articulation des mots fréquents en raison de réductions phonologiques dans l'histoire de la langue (Bybee, 2009, par exemple). En raison de cette corrélation, Fenk-Oczlon (1989) estime, par exemple, que le principe *court avant long* peut être substitué par le principe *très fréquent avant moins fréquent*. Les observations que nous avons faites à propos

12. La courbe est la représentation graphique du modèle de régression logistique suivant :

	Estimation	Erreur-type	valeur z	Pr(>|z|)
(Intercept)	-2.136	0.0303	-70.49	<2e-16
freq	-0.006	0.0001	51.06	<2e-16

des variables `freq` et `longAbs` montrent que les deux variables suivent les même tendances générales. Il reste à examiner plus précisément leur relation dans nos données. Nous évaluons la corrélation entre ces deux variables au moyen du ρ_s de Spearman, car, comme nous l'avons mentionné auparavant, cette mesure ne présuppose pas la linéarité de la relation entre les deux variables. Dans la table de données, $\rho_s = -0.50$ ($p < 2.2 \times 10^{-16}$), ce qui indique qu'il existe une relation inverse entre les deux variables et que la corrélation est importante. Néanmoins, la corrélation n'est pas forte au point que l'on puisse remplacer la longueur par la fréquence (ou inversement). La figure 4.3 donne une idée graphique de la corrélation entre les deux variables. Ce graphique a été obtenu en groupant les données autour de 21 groupes définis en fonction de la longueur de l'adjectif et en reportant pour chaque groupe la fréquence moyenne [13].

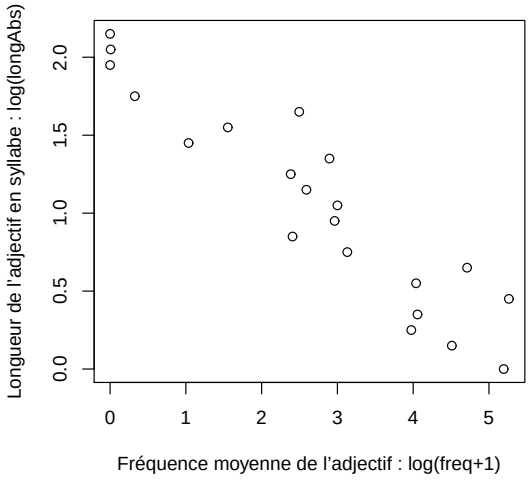

FIGURE 4.3.: Relation entre longueur et fréquence de l'adjectif dans nos données. Les données sont groupées autour de 21 intervalles définis selon la longueur de l'adjectif ; la fréquence correspond à la moyenne pour chaque intervalle.

4.2.3. Morphologie

De façon générale, les adjectifs morphologiquement construits tendent à être postposés. Afin de capter les effets de la morphologie dans nos données, nous avons distingué les adjectifs construits des adjectifs simples. En utilisant l'expression adjectifs construits, nous faisons référence aux adjectifs dérivés (*trompeur*, *substantiel*),

13. Nous avons utilisé les logarithmes de `freq` et `longAbs` pour le calcul du coefficient de corrélation et pour produire le graphique, car, en réduisant la dispersion des données, les valeurs logarithmiques permettent de rendre la corrélation plus claire.

aux participes présents (*suffisant*) et passifs (*vendu*) ainsi qu'aux adjectifs composés (*agro-alimentaire*).

L'annotation de nos données s'est faite en deux phases. Dans un premier temps, nous avons repéré automatiquement les adjectifs dérivés à l'aide de l'analyseur morphologique dérivationnel du français DériF (Namer, 2002). Cependant, cet outil ne couvre pas la totalité des règles de dérivation permettant de former des adjectifs. À titre d'exemple, le suffixe *-al* qui permet de dériver un adjectif à partir d'une base nominale, comme dans *artisan/artisanal*, *commune/communal*, n'est pas pris en compte par DériF. Nous avons donc, dans un deuxième temps, codé manuellement les adjectifs restants [14].

Dans la table de données, la morphologie des adjectifs est repérée par la variable binaire `construit`. Cette dernière prend la valeur 0 pour les adjectifs simples et 1 pour tous les autres adjectifs. Dans nos données, les adjectifs simples représentent seulement 497 lemmes, mais sont fréquents en termes d'occurrences. Inversement, les adjectifs construits renvoient à plus de lemmes distincts (1253) qui ont une fréquence moindre. Comme attendu, les adjectifs construits sont massivement postposés (93.7%), alors que les simples favorisent plutôt l'antéposition. La table 4.10 résume ces données.

	Antéposés		Postposés		Totaux		Nombre de lemmes
`construit` = 0	3458	50.7%	3363	49.3%	6821	100%	497
`construit` = 1	446	6.3%	6666	93.7%	7112	100%	1253

Table 4.10.: La variable `construit` en fonction de `position`

Pour avoir une meilleure image du comportement des adjectifs construits, nous avons distingué trois sous-classes : les composés, les participes et les adjectifs présentant le préfixe privatif *in-*. Pour les composés comme pour les participes, la préférence pour la postposition est massive (respectivement 97.4% et 91.6%). Plus précisément, seuls quatre composés apparaissent en antéposition : *extraordinaire*, *avant-dernier*, *sacro-saint* et *omnipuissant*. En ce qui concerne les participes, plus de 98.3% des participes passés sont postposés et on observe 18.8% d'antéposition pour les participes présents, avec des adjectifs comme *impressionnant* ou *angoissant*. Enfin, les adjectifs préfixés en *in-* ne semblent pas se comporter pas selon l'hypothèse que nous avions émise au chapitre précédent, à savoir que ces adjectifs favoriseraient l'antéposition. Ils préfèrent la postposition à 84.7%. Notons que ces adjectifs ne représentent que 196 occurrences, c'est-à-dire 1.4% des données de la table.

14. Dans une série de cas problématiques, nous avons fait des choix qui pourraient être discutés. Nous avons notamment considéré les adjectifs tels que *éminent* ou *présent* comme des adjectifs simples. En ce qui concerne les adjectifs se terminant par le suffixe *-iste* pour lesquels il existe un nom correspondant en *-isme*, nous avons décidé que l'adjectif était construit par dérivation à partir du nom. Enfin, pour un ensemble d'adjectifs de nationalité, tels que *anglais*, *allemand* et *catalan*, nous avons opté pour l'étiquette "simple". Nous tenons à remercier Delphine Tribout pour avoir pris le temps de valider notre classement morphologique. Toute erreur restante relève, cependant, de notre responsabilité.

Dans le but de vérifier si la distinction entre composés, participes et adjectifs préfixés en *in-* permet de mieux décrire le comportement de la variable à prédire `position`, nous construisons une nouvelle variable nommée `morpho`. Cette variable prend cinq valeurs distinctes : `non-construit`, `composé`, `participe`, `privatif` et `autres` construits [15].

Nous construisons un modèle de régression logistique dans lequel on évalue la probabilité d'antéposition de l'adjectif en fonction de la variable `morpho`. Le modèle, présenté dans la table 4.11, montre que les valeurs `participe`, `priv`, `composé` et `non_constr` ont un effet significatif sur la position de l'adjectif. La valeur de l'intercept capte l'effet de la variable `morpho` quand aucune valeur n'est associée à 1 (cf. tableau de la note 15), c'est-à-dire lorsque l'adjectif est un mot construit, mais pas un mot composé, un participe ou un dérivé en *in-* (valeur `autres` de la variable `morpho`). Ainsi, comme en atteste l'intercept négatif, ces adjectifs ont une forte préférence pour la postposition. Les autres coefficients associés aux valeurs de `morpho` dans le modèle doivent être interprétés en fonction de l'intercept. Par rapport à un adjectif construit (valeur `autres` de `morpho`), les adjectifs composés favorisent plus encore la postposition (coefficient négatif), alors que les participes et les dérivés en *in-* ont une préférence moins marquée pour cette position (coefficients positifs plus faibles que l'intercept). Enfin, les adjectifs non-construits ont une préférence plus forte que les autres pour l'antéposition comme en témoigne le coefficient positif élevé. Le modèle indique donc que la distinction des différents types de construction morphologique est pertinente pour décrire la position de l'adjectif.

	Estimation	Erreur-type	valeur z	Pr(>|z|)
(Intercept)	-2.78672	0.05715	-48.758	< 2e-16
morpho=part	0.39885	0.12877	3.097	0.00195
morpho=priv	1.07593	0.20646	5.211	1.87e-07
morpho=compos	-0.82420	0.34261	-2.406	0.01614
morpho=non_constr	2.75886	0.06207	44.445	< 2e-16

TABLE 4.11.: Paramètres du modèle de régression logistique avec `position` comme variable à prédire et `morpho` comme variable prédictrice

Le caractère morphologiquement construit des adjectifs est un aspect important

15. Techniquement, nous avons utilisé le codage disjonctif (*dummy coding*) pour coder la variable nominale `morpho`. Le codage est réalisé en utilisant la matrice suivante :

	part	priv	compos	non_constr
autres	0	0	0	0
part	1	0	0	0
priv	0	1	0	0
compos	0	0	1	0
non_constr	0	0	0	1

D'autres codages sont possibles pour les variables catégoriques, notamment le contraste d'Helmert (Venables & Ripley, 1999). Ce codage permet de mettre en lumière des contrastes entre les valeurs de la variable. Dans le cas de `morpho`, il nous a semblé suffisant d'utiliser le codage disjonctif.

dans la mesure où il touche un grand nombre de lemmes et qu'il favorise très fortement la postposition. Parmi les construits, les différences entre participes, adjectifs composés et adjectifs préfixés en *in-* ont un effet significatif sur le comportement de la variable `position`, ce qui indique que le type de construction morphologique a une influence sur la position de l'adjectif.

4.2.4. Classes lexicales

Les classes lexicales d'adjectifs constituent des facteurs importants dans le choix de la position de l'adjectif. Certaines classes semblent favoriser l'antéposition, comme les intensionnels et les évaluatifs, tandis que d'autres renforcent la postposition, comme la classe des adjectifs dénotant une catégorie objective.

Dans la table de données, nous avons annoté cinq catégories susceptibles d'avoir une influence sur le choix de la position de l'adjectif. Les catégories des relationnels et des subsectifs n'ont pas été prises en compte. Comme nous l'avons exposé dans la partie 3.8.1 (chapitre précédent), les relationnels ne forment pas une classe lexicale mais une classe d'emploi. À cause de la taille de notre table de données, il est apparu impossible d'annoter toutes les occurrences relationnelles. De plus, étant donné que les relationnels sont, dans une très large majorité des dénominaux, la variable `construit` capte en partie les adjectifs potentiellement relationnels. En ce qui concerne la classe des subsectifs, elle semble difficilement opératoire pour l'annotation. Cette classe renvoie à « *une qualité relative, qui dépend de la catégorie référentielle à laquelle elle s'applique* » (Noailly, à paraître). Ainsi, l'adjectif *petit* est subsectif car une *petite maison* est plus grande qu'un *petit vélo*. Les adjectifs de couleur ne sont normalement pas considérés comme des subsectifs. Pourtant, quand on parle d'une *peau noire*, la qualité du noir n'est pas la même que lorsqu'on parle d'*encre noire*. En raison de la difficulté à tracer les contours de cette classe, nous l'avons laissée de côté.

Les adjectifs de nationalité et les adjectifs de couleur ont été repérés automatiquement à l'aide des dictionnaires PROLEXBASE (Tran & Maurel, 2006) et CHROMA [16]. Les adjectifs intensionnels [17] et les adjectifs évaluatifs ont été annotés manuellement. Enfin, les adjectifs indéfinis sont au nombre de 6 : *différent*, *autre*, *certain*, *quelconque*, *divers*, *tel*. Les variables codant ces classes sont les suivantes :

`couleur`

= 1 : l'adjectif est un adjectif de couleur,
= 0 : l'adjectif n'est pas un adjectif de couleur ;

`natio`

= 1 : l'adjectif est un adjectif de nationalité,
= 0 : l'adjectif n'est pas un adjectif de nationalité ;

16. http://pourpre.com/chroma/
17. La liste des intensionnels est : *actuel*, *ancien1*, *apparent*, *authentique*, *dit*, *éventuel*, *faux*, *futur*, *imminent*, *nouveau*, *parfait*, *passé*, *possible*, *potentiel*, *précédent*, *prétendu*, *probable*, *prochain*, *promis*, *réel*, *temporaire*, *théorique*, *véritable*, *virtuel*, *vrai*.

intens

= 1 : l'adjectif est un adjectif intensionnel,
= 0 : l'adjectif n'est pas un adjectif intensionnel ;

eval

= 1 : l'adjectif est un adjectif évaluatif,
= 0 : l'adjectif n'est pas un adjectif évaluatif ;

indef

= 1 : l'adjectif est un adjectif indéfini,
= 0 : l'adjectif n'est pas un adjectif indéfini ;

Les proportions relatives à ces variables sont présentées dans la table 4.12.

	Antéposés		Postposés		Totaux		Nombre de lemmes
Données générales	3809	27.3%	10124	72.7%	13933	100%	1750
`couleur` =1	0	0%	61	100%	61	100%	21
`natio` = 1	1	0.1%	1796	99.9%	1797	100%	138
`eval` = 1	360	68.6%	165	31.4%	525	100%	59
`intens` = 1	736	67.3%	358	32.7%	1094	100%	25
`indef` = 1	376	90.6%	39	9.4%	415	100%	6

TABLE 4.12.: Les variables relatives aux classes lexicales en fonction de `position`

Les adjectifs de nationalité et de couleur sont tous postposés, excepté un adjectif de nationalité qui apparaît en antéposition dans la phrase (8).

(8) *Signe des temps : la très* ***britannique*** *banque d'affaires et de marché vient d'acheter un siège à la Bourse de Paris.*

Les indéfinis se présentent massivement en antéposition. Les deux dernières classes ont une préférence marquée pour l'antéposition également.

4.2.5. Syntaxe

Les contraintes syntaxiques interviennent à deux niveaux. Les plus importantes agissent au sein du SADJlorsque l'adjectif est coordonné ou modifié. Les autres se situent au niveau du SN et concernent les autres constituants présents, le déterminant introduisant le nom ainsi que la fonction du SN.

L'ensemble des variables syntaxiques a été automatiquement extrait à partir de l'annotation en constituants et en fonctions du FTB. Les deux variables concernant la configuration du SADJ sont `coord` et `adv`.

coord

= 1 : l'adjectif apparaît dans une coordination,
= 0 : l'adjectif n'est pas coordonné ;

adv

= 1 : l'adjectif est modifié par un adverbe ou une locution adverbiale en position pré-adjectivale,
= 0 : l'adjectif n'est pas modifié ;

Lorsque ces deux variables ont la valeur 1, le SAdj contient un ou plusieurs mots en plus, ce qui signifie que le SAdj est plus long. D'après le critère de longueur, vu dans la section 4.2.1, ces deux variables favorisent la postposition. Cependant, alors que la variable `coord` favorise très fortement la postposition, les adjectifs associés à un adverbe ne sont que légèrement sur-représentés en postposition, comme le montrent les proportions de la table 4.13.

	Antéposés		Postposés		Totaux	
Données générales	3809	27.3%	10124	72.7%	13933	100%
`coord` =1	43	5.7%	708	94.3%	751	100%
`adv` = 1	192	25.6%	559	74.4%	751	100%

Table 4.13.: Les variables concernant le SAdj en fonction de `position`

Les données confirment l'hypothèse formulée par Abeillé & Godard (1999), à savoir que la position antéposée n'admet qu'un paradigme restreint d'adverbes. Dans nos données, seuls onze lemmes adverbiaux se rencontrent en antéposition [18], alors qu'on dénombre, en postposition, 120 adverbes et locutions adverbiales différents. Cependant, les données montrent que les adverbes de degré ne sont pas les seuls à pouvoir apparaître en antéposition. Nous donnons ici l'exemple d'un SAdj antéposé contenant le modifieur *désormais*.

(9) *…avant de se consoler un peu avec le* ***désormais*** *traditionnel petit rallye de fin décembre qui lui permet de terminer l'année sur une progression de 5,22%.*

Les contraintes relatives à la configuration interne du SN sont captées à l'aide de quatre variables codant la présence ou l'absence d'autres constituants [19].

adjAnt

= 1 : il y a (au moins) un adjectif antéposé dans le SN,
= 0 : il n'y a pas d'adjectif antéposé dans le SN ;

adjPost

= 1 : il y a (au moins) un adjectif postposé dans le SN,
= 0 : il n'y a pas d'adjectif postposé dans le SN ;

18. Les onze adverbes modifiant un adjectif antéposé sont : *trop*, *très*, *tout*, *si*, *plus*, *peu*, *moins*, *encore*, *désormais*, *bien*, *aussi*.

19. Nous n'avons pas pris en compte la présence de suborbonnées (Ssub) et d'infinitives (VPinf) dans le SN car elles représentaient très peu de données.

relative

= 1 : il y a (au moins) une relative dans le SN,
= 0 : il n'y a pas de relative dans le SN ;

sprep

= 1 : il y a (au moins) un SP dans le SN,
= 0 : il n'y a pas de SP dans le SN ;

Si l'on suit l'idée selon laquelle on a tendance à équilibrer le nombre d'éléments de part et d'autre du nom tête, les variables `adjPost`, `relative` et `sprep` doivent favoriser l'antéposition. Inversement, la présence d'un adjectif en antéposition (`adjAnt` = 1) devrait favoriser la postposition d'un nouvel adjectif. Les données, résumées dans la table 4.14, sont cohérentes avec ces attentes. La proportion d'antéposition est nettement supérieure à celle des données générales, dans le cas où le SN contient un SP (`sprep` = 1). Les préférences respectives de `adjAnt` et `adjPost` sont moins marquées que celle de la variable `sprep`. Enfin, en ce qui concerne la variable `relative` la tendance est très faible et n'est pas statistiquement significative ($\chi^2(1) = 1.0287$, p= 0.31).

	Antéposés		Postposés		Totaux	
Données générales	3809	27.3%	10124	72.7%	13933	100%
`adjAnt` =1	187	22.9%	628	77.1%	815	100%
`adjPost` = 1	537	32.0%	1139	68.0%	1676	100%
`relative` = 1	170	29.3%	411	70.7%	581	100%
`sprep` = 1	1491	38.9%	2345	61.1%	3836	100%

TABLE 4.14.: Les variables relatives à la configuration du SN en fonction de `position`

D'après Forsgren (1978), la nature du déterminant introduisant le SN est un indice formel permettant de capter en partie le comportement de la position de l'adjectif. Pour observer l'effet du déterminant, nous disposons des quatre variables suivantes :

absDet

= 1 : le SN n'est pas introduit par un déterminant,
= 0 : le SN est introduit par un déterminant ;

detPoss

= 1 : le SN est introduit par un déterminant possessif,
= 0 : le SN n'est pas introduit par un déterminant possessif ;

detDem

= 1 : le SN est introduit par un déterminant démonstratif,
= 0 : le SN n'est pas introduit par un déterminant démonstratif ;

artDef

= 1 : le SN est introduit par un article défini,
= 0 : le SN n'est pas introduit par un article défini.

Globalement, nos données sont en accord avec celles de Forsgren, dans la mesure où l'on observe une légère préférence pour l'antéposition lorsque le déterminant est défini (`detPoss` = 1 ou `artDef` = 1 ou `detDem` = 1). L'observation plus précise de la nature du déterminant montre que le déterminant démonstratif présente la proportion la plus importante d'adjectifs antéposés (54.9%), suivi du déterminant possessif (33.5%). En revanche, l'article défini semble légèrement favoriser la postposition avec 76.5% d'adjectifs postposés. Enfin, dans le cas de la variable `absDet`, la postposition semble être préférée, mais la tendance est très légère. Ces chiffres sont récapitulés dans la table 4.15.

	Antéposés		Postposés		Totaux	
Données générales	3809	27.3%	10124	72.7%	13933	100%
`artDef` =1	1664	23.5%	5412	76.5%	815	100%
`detPoss` = 1	187	33.5%	372	66.5%	1676	100%
`detDem` = 1	164	45.1%	200	54.9%	581	100%
`absDet` = 1	565	25.2%	1675	74.8%	3836	100%

TABLE 4.15.: Les variables relatives au déterminant introduisant le SN en fonction de `position`

En référence au travail de Forsgren (1978), nous avons également annoté la table de données selon la fonction que remplit le SN contenant l'adjectif étudié. Nous nous sommes appuyée sur l'annotation fonctionnelle du FTB et nous avons retenu trois fonctions réservées aux SN (sujet, objet direct, attribut du sujet) et deux fonctions qui peuvent être remplies par un SP (modifieur et objet non-direct [20]). Dans le cas du modifieur et de l'objet non-direct, le SN ne remplit pas directement la fonction, mais il est le complément de la préposition introduisant le SP qui assume la fonction.

`sujet`

= 1 : le SN a la fonction sujet,
= 0 : le SN n'a pas la fonction sujet ;

`objet`

= 1 : le SN a la fonction objet,
= 0 : le SN n'a pas la fonction objet ;

`ats`

= 1 : le SN a la fonction attribut du sujet,
= 0 : le SN n'a pas la fonction attribut du sujet ;

`objNonDirect`

= 1 : le SN appartient à un SP ayant la fonction objet non-direct,
= 0 : le SN n'appartient pas à un SP ayant la fonction objet non-direct ;

20. Nous utilisons l'étiquette objet non-direct pour regrouper les fonctions A_OBJ, DE_OBJ et P_OBJ du FTB.

`modifieur`

= 1 : le SN ou le SP dans lequel apparaît le SN a la fonction modifieur,
= 0 : le SN ou le SP dans lequel apparaît le SN n'a pas la fonction modifieur ;

Contrairement à ce qu'observe Forsgren (1978), la variable `sujet` ne semble pas pertinente car la proportion de postposition quand le SN est sujet, est identique à celle des données générales. De même, la fonction `objNonDirect` ne paraît pas favoriser une position par rapport à l'autre. En revanche, les variables `modifieur`, `ats` et `objet` montrent une légère préférence pour l'antéposition. Les proportions exactes sont présentées dans la table 4.16.

	Antéposés		Postposés		Totaux	
Données générales	3809	27.3%	10124	72.7%	13933	100%
`sujet` =1	595	27.3%	1582	72.7%	2177	100%
`objet` = 1	684	31.0%	1524	69.0%	2208	100%
`ats` = 1	156	41.7%	218	58.3%	374	100%
`objNonDirect` = 1	256	26.9%	694	73.1%	950	100%
`modifieur` = 1	815	33.0%	1658	67.0%	2473	100%

TABLE 4.16.: La variable `position` selon les variables relatives à la fonction du SN

Il est particulièrement difficile d'interpréter les proportions relatives au déterminant et à la fonction du SN dans la mesure où, si elles ont un réel effet, ces variables sont secondaires par rapport aux contraintes lexicales ou celles concernant la configuration du SAdj (`adv` et `coord`). Il est donc essentiel de les étudier dans un modèle où les contraintes ayant un effet plus massif sont prises en compte.

4.2.6. Combinaison du nom et de l'adjectif

Le choix de la position d'un adjectif peut être influencé par le nom avec lequel il se combine. C'est notamment le cas des séquences collocatives. Nous entendons collocation dans un sens large, tel qu'il est défini par Manning & Schütze (1999), par exemple : « *une collocation est une expression composée de deux ou plusieurs mots qui correspond à une façon conventionnelle de dire les choses* » [21].

Les collocations impliquant un nom et un adjectif ont un effet sur la position de l'adjectif, car l'une des conventions imposée par l'usage est l'ordre dans lequel apparaissent les deux mots. Ainsi, certains noms ont tendance à se combiner avec un adjectif intensificateur spécifique, comme *lourd tribut* ou *vibrant hommage*. La spécificité de la combinaison ne tient pas seulement au choix de l'adjectif intensificateur, elle tient également à l'ordre de la séquence. Par exemple, l'expression collocative *vibrant hommage* contient un adjectif dont les propriétés, telles que sa nature morphologique, favorisent la postposition (comme dans *voix vibrante*, *ton vibrant*). Or, la présence

21. « *a collocation is an expression consisting of two or more words that correspond to some conventional way of saying things* », (Manning & Schütze, 1999, p. 151)

du nom *hommage* favorise fortement son antéposition en raison de la fréquence de l'apparition de ces deux mots dans cet ordre-là. Pour d'autres séquences, l'ordre des éléments est aussi lié à la sémantique. À titre d'exemple, la séquence collocative *(à) juste titre* n'a pas le même sens que *titre juste*.

La prise en compte de l'adjectif en fonction du nom et de sa position par rapport à ce nom semble donc pertinente du point de vue de l'usage et de la sémantique. De plus, nous estimons que, dans le cadre de notre travail, la plupart des expressions figées telles que celles évoquées dans la section 3.6 du chapitre 3, sont détectables à partir des collocations nom-adjectif. Par exemple, pour détecter les expressions *à bras raccourcis* et *en chute libre*, il suffit de repérer la co-occurrence des items nominaux et adjectivaux.

Pour obtenir une estimation statistique de la relation d'association entre l'item adjectival et l'item nominal, nous avons employé une métrique couramment utilisée pour identifier les collocations : le χ^2 (Manning & Schütze, 1999). Le calcul des collocations nécessite un gros volume de données, c'est pourquoi nous avons eu recours au corpus ER. Nous avons procédé à l'extraction des bigrammes de lemmes Adjectif - Nom et Nom - Adjectif et créé 2 listes, l'une servant à identifier les collocations avec adjectif antéposé, l'autre pour les collocations avec adjectif postposé. À partir de ces listes, nous avons calculé la valeur de χ^2 de chaque bigramme. Dans la table de données, nous avons introduit deux variables, `collocAN` et `collocNA`, qui donnent les scores d'association de chaque couple Nom - Adjectif[22]. Afin de réduire l'étendue des valeurs de ces variables et ainsi réduire l'effet des valeurs extrêmes, nous avons effectué une transformation logarithmique. Plus précisément, les variables `collocAN` et `collocNA` ont pour valeur $log(\chi^2 + 1)$, afin d'éviter d'avoir des valeurs infinies ($= log(0)$) dans le cas des couples Nom - Adjectif qui n'ont pas été rencontrés dans le corpus ER.

`collocAN` : score de χ^2 pour la séquence ordonnée Adjectif - Nom (échelle logarithmique) ;

`collocNA` : score de χ^2 pour la séquence ordonnée Nom - Adjectif (échelle logarithmique).

Le rôle de ces variables est de donner une estimation de la force de l'association de l'adjectif par rapport au nom. Les 20 séquences Adjectif - Nom et Nom - Adjectif avec le score de χ^2 le plus élevé sont représentées respectivement dans le graphique de la figure 4.4 et dans celui de la figure 4.5.

Pour évaluer l'effet de ces variables, nous observons les proportions d'antéposition et de postposition selon que les variables sont égales à 0 ou supérieures à 0. La table 4.17 montre qu'un score supérieur à 0 favorise l'antéposition dans le cas de `collocAN` et la postposition dans le cas de `collocNA`.

22. Nous avons modifié une valeur de χ^2 manuellement. Le bigramme *chaude eau* a obtenu le cinquième score de χ^2 pour les séquences Adjectif - Nom. Or toutes les occurrences de la séquence *chaude eau* font référence à un quartier d'une commune ("La Chaude Eau"), fréquemment cité dans ER. Étant donné que cette séquence n'est pas attestée dans le corpus ER dans un autre contexte, nous avons décidé de donner la valeur 0 à la variable `collocAN` pour le couple *eau/chaud*.

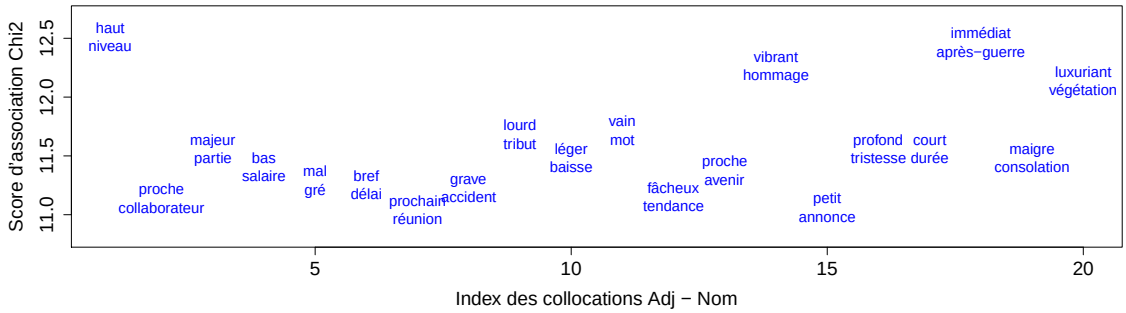

FIGURE 4.4.: Collocations Adjectif - Nom ayant le score le plus élevé.

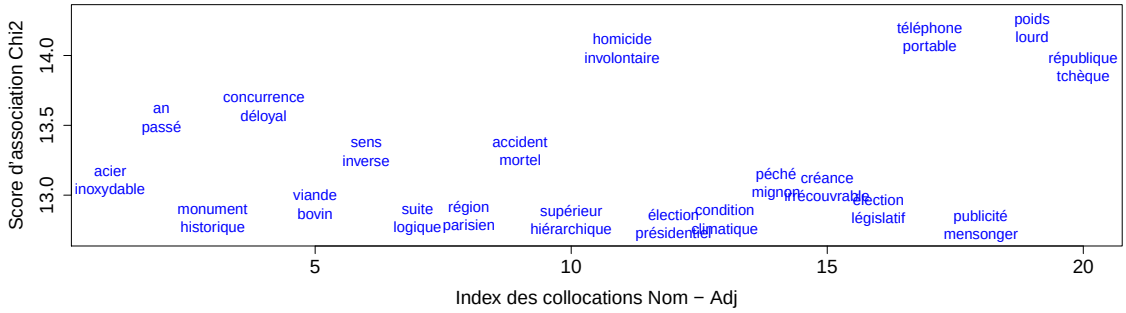

FIGURE 4.5.: Collocations Nom - Adjectif ayant le score le plus élevé.

Ces données montrent clairement que prendre en compte le couple Nom - Adjectif est un moyen de déterminer la position de l'adjectif de façon massive. En repérant les couples Nom - Adjectif avec un fort degré d'association, ces deux variables permettent de prendre en compte une partie des spécificités sémantiques pouvant intervenir dans la combinaison d'un nom et d'un adjectif, comme nous l'avons évoqué avec *juste titre*.

4.2.7. Liaison et hiatus

Tout d'abord, il faut préciser que nous travaillons sur des données écrites. Si les contraintes concernant la liaison et le hiatus ont réellement un effet sur le choix de la position, cet effet est très probablement amoindri par le mode de production des données. Néanmoins, il est intéressant de contrôler si ce type de contraintes est observable dans nos données. De plus, ces contraintes n'ont pas d'effet massif sur le choix de la position. Ainsi, pour avoir une véritable idée de leur rôle respectif, il faut les intégrer à un modèle prenant en compte l'ensemble des contraintes, ce que nous ferons dans la partie 4.3.

Dans le chapitre précédent, nous avons émis l'hypothèse qu'une contrainte préférentielle anti-hiatus pouvait influencer la position de l'adjectif : lorsqu'il existe un

	Antéposés		Postposés		Totaux	
`collocAN` = 0	777	7.9%	9071	92.1%	9848	100%
`collocAN` > 0	3032	74.2%	1053	25.8%	4085	100%
`collocNA` = 0	2432	42.5%	3293	57.5%	5725	100%
`collocNA` > 0	1377	16.8%	6831	83.2%	8208	100%

TABLE 4.17.: Les variables `collocAN` et `collocNA` en fonction de `position`

hiatus potentiel en postposition, l'adjectif a tendance à être antéposé pour éviter le hiatus. Pour étudier l'effet de cette contrainte, nous utilisons la variable `hiatusPost`. Cette variable est binaire et s'interprète de la façon suivante :

`hiatusPost`

= 1 : il existe un hiatus potentiel en postposition,
= 0 : il n'existe pas de hiatus potentiel en postposition.

Pour obtenir cette variable, nous avons utilisé *Lexique 3.72* (New, 2006; New *et al.*, 2001). En nous appuyant sur la transcription phonologique contenue dans cette ressource, nous avons défini les lemmes adjectivaux commençant par une voyelle et les lemmes nominaux finissant par une voyelle. Rappelons que quand le SN est au pluriel, la liaison est toujours possible pour éviter le hiatus. Nous limitons donc les observations concernant le hiatus aux SN singuliers. La variable `hiatusPost` prend la valeur 1 lorsque, dans un SN singulier, le nom finit par une voyelle et que l'adjectif commence par une voyelle.

Cette variable `hiatusPost` concerne 1084 occurrences et l'effet observé est inverse à celui attendu. Comme le montre le tableau 4.18, l'existence d'un hiatus potentiel en postposition tend à favoriser cette position, avec 85.2% d'adjectifs postposés. Cependant, ces proportions sont à prendre avec précaution, comme nous le verrons dans la partie modélisation.

	Antéposés		Postposés		Totaux	
`hiatusPost`= 1	160	14.8%	924	85.2%	1084	100%

TABLE 4.18.: La variable `hiatusPost` en fonction de `position`

La deuxième hypothèse relative aux problèmes de liaison concerne la forme de liaison de masculin singulier (FLMS) et plus particulièrement les adjectifs défectifs pour cette forme (Bonami & Boyé, 2003; Morin, 1992), par exemple *franc*, *blond*. L'idée est que les adjectifs défectifs pour la FMLS ont tendance à être postposés dans un contexte masculin singulier.

L'inventaire de ces adjectifs est difficile à dresser dans la mesure où il existe souvent une hésitation entre la liaison et le hiatus pour les adjectifs à consonne latente, mais il est difficile de statuer sur le fait que ces adjectifs sont réellement défectifs. Par exemple, la production de la séquence *un haut immeuble* semble provoquer une hésitation entre une hiatus et une liaison en [t]. Faut-il alors considérer que cet adjectif

est défectif pour la FLMS ? Il est difficile d'en juger à partir de ce type de données. Nous pensons que le seul moyen de pouvoir réellement faire l'inventaire des adjectifs défectifs pour la FLMS est de mettre en place une expérience où les locuteurs doivent oraliser de telles séquences et observer les productions effectives. Une telle entreprise ne rentrant pas dans le cadre de cette thèse, nous avons testé une contrainte plus générale qui pourrait être liée à la question de la défectivité pour la FLMS. Cette contrainte concerne les adjectifs à consonne latente, à savoir les adjectifs présentant une forme à finale vocalique au masculin et à consonne finale au féminin, par exemple *précis/précise* ou *sain/saine*[23]. L'idée est que la présence de la consonne latente peut provoquer l'hésitation entre le hiatus et la liaison en antéposition, et donc favoriser la postposition en cas de SN masculin singulier.

En utilisant la transcription des formes du masculin et du féminin de *Lexique 3.72*, nous avons repéré les adjectifs à consonne latente. Premièrement, nous avons éliminé de cet inventaire sept adjectifs qui ont un comportement particulier : leur forme de liaison ne correspond pas à la forme du féminin, mais à une autre forme, comme dans *un grand-t-appartement* ou *un gros-z-avion*. Ces sept adjectifs sont *grand*, *profond*, *second*, *bas*, *doux*, *faux* et *gros* (inventaire tiré de Bonami & Boyé, 2005, p. 81). Deuxièmement, nous avons écarté de cet inventaire une série d'adjectifs pour lesquels la production de la liaison ne pose aucun problème selon notre jugement : *petit*, *dernier*, *premier*, *certain*, *bon*, *prochain* , *ancien*, *mauvais*. Troisièmement, nous avons supprimé les adjectifs de nationalité de cet inventaire. En effet, un nombre important d'adjectifs de nationalité ont une consonne latente (*français*, *européen*...), mais ces adjectifs sont très largement postposés et ce pour des raisons autres que des problèmes de liaison. Nous avons donc fait le choix de les écarter. Nous désignerons par l'expression classe restreinte à consonne latente, la classe d'adjectifs ayant une consonne latente et n'appartenant pas à l'un des trois cas répertoriés ci-dessus.

La variable qui nous intéresse s'appelle `conLatMS`. Elle est binaire et permet de coder les contextes de potentielle hésitation entre la liaison et le hiatus.

`conLatMS`

= 1 : contexte avec un nom à initiale vocalique et un adjectif masculin singulier appartenant à la classe restreinte à consonne latente,
= 0 : tous les autres contextes.

Si l'hypothèse selon laquelle l'hésitation entre le hiatus et la liaison favorise la postposition est vraie, la proportion de postposition doit être plus élevée lorsque `conLatMS` est égale à 1. Les données dont nous disposons sont limitées, car seules 92 occurrences répondent positivement au contexte pouvant provoquer une hésitation. Les observations sont cohérentes avec la tendance attendue, comme le montre la table 4.19. Cependant, le nombre de données pertinentes étant trop restreint, la différence de proportion n'est significative qu'au seuil 0.05 : $\chi^2(1) = 3.83$ (p= 0.05).

23. Nous incluons les adjectifs présentant une alternance entre une voyelle nasale au masculin et la voyelle équivalente non-nasale suivie d'un [n] au féminin.

	Antéposés		Postposés		Totaux	
conLatMS=1	17	17.9%	78	82.1%	95	100%
conLatMS=0	3792	27.4%	10046	72.6%	13838	100%

TABLE 4.19.: La variable conLatMS en fonction de position

4.2.8. Bilan

Nous avons décrit l'ensemble des variables prises en compte dans notre table de données. Notons que les variables concernant la fréquence, la longueur, les aspects lexicaux et la combinaison Nom - Adjectif semblent avoir une réelle influence sur le choix de la position de l'adjectif. Les variables syntaxiques et celles relatives aux problèmes de liaison ont, à première vue, moins d'importance. Nous récapitulons l'ensemble des variables dans la liste suivante :

longAbs nombre de syllabes de l'adjectif

ratioLong longSAdj/longAbs

freq fréquence de l'adjectif dans ER

morpho l'adjectif est un adjectif composé, un adjectif préfixé en *in-*, un participe, un autre type d'adjectif construit ou un adjectif non-construit

couleur l'adjectif est un adjectif de couleur ou non

natio l'adjectif est un adjectif de nationalité ou non

intens l'adjectif est un adjectif intensionnel ou non

eval l'adjectif est un adjectif évaluatif ou non

indef l'adjectif est un adjectif indéfini ou non

coord l'adjectif apparaît dans une coordination ou non

adv l'adjectif est modifié ou non

adjAnt il y a (au moins) un adjectif antéposé dans le SN ou non

adjPost il y a (au moins) un adjectif postposé dans le SN ou non

relative il y a (au moins) une relative dans le SN ou non

sprep il y a (au moins) un SP dans le SN ou non

absDet le SN est introduit par un déterminant ou non

detPoss le SN est introduit par un déterminant possessif ou non

detDem le SN est introduit par un déterminant démonstratif ou non

artDef le SN est introduit par un article défini ou non

sujet le SN a la fonction sujet ou non

objet le SN a la fonction objet ou non

ats le SN a la fonction attribut du sujet ou non

objNonDirect le SN appartient à un SP ayant la fonction objet non-direct ou non

modifieur le SN ou le SP dans lequel apparaît le SN a la fonction modifieur ou non

collocAN score de χ^2 pour la séquence ordonnée Adjectif - Nom (échelle logarithmique)

collocNA score de χ^2 pour la séquence ordonnée Nom - Adjectif (échelle logarithmique)

hiatusPost il existe un hiatus potentiel en postposition ou non

conLatMS contexte avec un nom à initiale vocalique et un adjectif masculin singulier appartenant à la classe restreinte à consonne latente ou non

4.3. Modèles

L'objectif de cette section est double. Premièrement, nous cherchons à modéliser le phénomène de l'alternance de position dans nos données à partir de l'ensemble des variables prédictrices dont nous disposons. Pour cela, nous utiliserons la régression logistique et les modèles à effets mixtes. Deuxièmement, nous cherchons à estimer l'effet des différents types de contraintes. La méthodologie déployée repose sur la comparaison de modèles construits sur différents faisceaux de contraintes [24].

Après avoir précisé quelques aspects techniques, nous présenterons trois modèles que nous appellerons Modèle Syntaxe, Modèle Collocation et Modèle Lexical. Nous introduirons ensuite un modèle à effets aléatoires permettant de rendre compte au mieux du comportement de chaque adjectif. Enfin, nous présenterons le Modèle Global. Issu de la compilation des modèles précédents, il représente la modélisation la plus aboutie du phénomène de l'alternance de position de l'adjectif.

4.3.1. Aspects "techniques"

Les techniques statistiques utilisées sont présentées dans le chapitre 2. Nous modélisons la probabilité d'antéposition de l'adjectif, autrement dit la probabilité que `position` = 1. Le modèle de régression logistique que nous utilisons se définit de la façon suivante :

$$P(\texttt{position} = 1|X) = \frac{e^{\beta X}}{1 + e^{\beta X}} \tag{4.1}$$

où X renvoie au vecteur contenant les variables prédictrices utilisées.

Nous construisons un Modèle Nul, qui ne contient aucune variable prédictrice et qui prédit systématiquement l'échec (`position` = 0), c'est-à-dire la postposition. Pour ce Modèle Nul, l'exactitude au seuil $P(\texttt{position} = 1|X) = 0.5$ est $E = 0.727$, ce

24. Cette méthodologie est également utilisée dans les articles sur lesquels s'appuie ce travail, notamment Thuilier *et al.* (2012) et Fox & Thuilier (2010).

qui correspond à la proportion d'adjectifs postposés dans la table de données. L'aire sous la courbe ROC (*AUC*), qui est une autre mesure de la qualité d'un modèle, est égale à 0.5.

Les modèles que nous allons présenter, ont été compactés sur la base du test de rapport de vraisemblance. Étant donné deux modèles imbriqués l'un dans l'autre, si le rapport de vraisemblance est statistiquement significatif, on considère que le modèle le plus complexe est justifié pour la modélisation de la variable `position`. En revanche, si le rapport de vraisemblance n'est pas significatif, on considère que le modèle le plus simple suffit à la modélisation de la variable `position`. Pour le modèle le plus compact, nous nous assurerons de la qualité de ce dernier en vérifiant le niveau de multicolinéarité (indice de conditionnement κ et facteur d'inflation de la variance), puis en calculant son exactitude et l'aire sous la courbe ROC. Le Modèle Nul que nous venons de présenter servira de point de référence pour ces mesures.

Les variables numériques `longAbs` et `freq` sont passées au logarithme pour réduire l'intervalle sur lequel s'étendent leurs valeurs et ainsi réduire l'effet des valeurs extrêmes. Plus exactement, pour `freq`, nous avons appliqué la transformation $log(\texttt{freq} + 1)$, car certaines fréquences sont égales à 0. De plus, ces deux variables sont centrées autour de 0, afin de réduire au maximum la colinéarité. Les variables `longAbs` et `freq` correspondent donc respectivement au logarithme de la longueur centré et au logarithme de la fréquence centré.

Enfin, comme nous l'avons vu dans la description des classes lexicales (cf. section 4.2.4), la variable `couleur` ne présente aucun cas d'antéposition dans nos données. Nous ne la prendrons pas en compte dans la modélisation car ce type de variable pose des problèmes dans l'estimation des paramètres des modèles.

4.3.2. Modèle Syntaxe

Le premier modèle concerne les aspects syntaxiques de l'alternance de position de l'adjectif. Nous regroupons les variables relatives à la configuration interne du SAdj (`coord`, `adv`), à la configuration du SN (`adjAnt`, `adjPost`, `sprep`, `relative`, `artDef`, `detPoss`, `detDem`, `absDet`) et à la fonction assumée par le SN (`sujet`, `objet`, `ats`, `modifieur`, `objNonDirect`). Nous avons également introduit la variable `ratioLong`. Dans la mesure où elle capte le fait que le SAdj contient des éléments autres que l'adjectif, cette variable rend compte d'un aspect de la syntaxe du SAdj. Nous avons construit un modèle de régression logistique à partir de ces 16 variables, puis nous l'avons compacté en utilisant le test du rapport de vraisemblance. Les variables `detPoss`, `objNonDirect`, `relative` et `ratioLong` ont été écartées par cette méthode. L'éviction de `ratioLong` tient au fait que cette dernière n'apporte pas plus d'informations que les variables `adv` et `coord`. Le modèle obtenu contient 12 variables. Il est présenté dans la table 4.20.

Dans ce modèle, les variables avec un coefficient positif votent pour l'antépostion, et celles ayant un coefficient négatif pour la postposition. Ainsi, les variables `detDem`, `adjPost`, `sprep`, `modifieur`, `ats`, `objet` et `sujet` favorisent l'antéposition, tandis que `artDef`, `absDet`, `coord`, `adjAnt` et `adv` préfèrent la postposition. Notons que toutes

	Estimation	Erreur-type	valeur z	Pr(>|z|)	
(Intercept)	-1.06552	0.04773	-22.322	< 2e-16	***
artDef=1	-0.52583	0.04645	-11.319	< 2e-16	***
detDem=1	0.64224	0.11481	5.594	2.22e-08	***
absDet=1	-0.31105	0.06365	-4.887	1.02e-06	***
coord=1	-1.84679	0.15985	-11.553	< 2e-16	***
adjAnt=1	-0.28289	0.08814	-3.210	0.00133	**
adjPost=1	0.34640	0.05831	5.940	2.84e-09	***
sprep=1	0.79789	0.04221	18.902	< 2e-16	***
adv=1	-0.27861	0.08929	-3.120	0.00181	**
modifieur=1	0.47158	0.05345	8.823	< 2e-16	***
ats=1	0.79271	0.11391	6.959	3.42e-12	***
objet=1	0.19105	0.05904	3.236	0.00121	**
sujet=1	0.18554	0.05951	3.118	0.00182	**

(Significativité des effets selon le test de Wald : *** p<0.001, ** p<0.01, * p<0.05)

TABLE 4.20.: Paramètres du Modèle Syntaxe

ces variables sont binaires, ce qui permet de comparer leurs coefficients directement. De plus, la multicolinéarité du modèle est très réduite ($\kappa = 5.359409$ [25]). On peut donc considérer que le coefficient affecté à chaque variable rend compte de la valeur explicative de la variable en question. Cependant, la qualité de prédiction du modèle n'est pas bonne. Lorsque l'on fixe un seuil de décision à $P(\texttt{position} = 1|X) = 0.5$, l'exactitude par validation croisée 100 passes est $\mu = 0.726$ ($\sigma = 0.04$), ce qui signifie que les capacités de prédiction de ce modèle ne sont pas supérieures à celles du Modèle Nul. La matrice de confusion du modèle est présentée dans le tableau 4.21. On observe que ce modèle permet une légère amélioration de la prédiction en antéposition par rapport au Modèle Nul, mais les performances en postposition sont légèrement moins bonnes.

		Prédits		%
		position=1	position=0	correct
Observés	position=1	219	3590	5.7%
	position=0	237	9887	97.7%
Exactitude				$\mu = 0.726$ ($\sigma = 0.04$)

TABLE 4.21.: Matrice de confusion du Modèle Syntaxe

25. Cette valeur de κ indique qu'il n'y a pas de multicolinéarité dans le modèle, ce qui est confirmé par les facteurs d'inflation de la variance (VIF) des variables du modèle qui sont tous compris entre 1 et 1.4.

L'aire sous la courbe ROC indique également que ce modèle ne permet pas une classification satisfaisante de la variable à prédire : $AUC = 0.66$. Ces résultats montrent que les variables liées à la syntaxe du SAdj et du SN ne permettent pas de produire une classification plus satisfaisante de la variable `position` que le Modèle Nul. Les contraintes syntaxiques ne sont donc pas pertinentes lorsqu'elles sont considérées de façon isolée.

4.3.3. Modèle Collocation

Le Modèle Collocation ne contient que deux variables : `collocAN` et `collocNA`. Il permet de mesurer l'importance, pour déterminer la position de l'adjectif, du nom avec lequel ce dernier est combiné. Le Modèle Collocation est présenté dans la table 4.22.

	Estimation	Erreur-type	valeur z	Pr(>|z|)	
(Intercept)	-1.21584	0.02963	-41.03	<2e-16	***
collocNA	-0.60248	0.01735	-34.72	<2e-16	***
collocAN	0.86876	0.01849	46.99	<2e-16	***

(Significativité des effets selon le test de Wald : *** p<0.001, ** p<0.01, * p<0.05)

Table 4.22.: Paramètres du Modèle Collocation

Comme attendu, la variable `collocAN` vote pour l'antéposition et `collocNA` pour la postposition. Plus exactement, le modèle prédit par défaut la postposition et une valeur supérieure à 0 pour `collocNA` renforce la prédiction de la postposition. Si la variable `collocNA` est égale à 0, une valeur de plus de 1.4 pour la variable `collocAN` permet au modèle de prédire l'antéposition. La multicolinéarité de ce modèle est très faible comme le prouvent l'indice de conditionnement et le facteur d'inflation de la variance : $kappa = 2.32828$ et $VIF(\texttt{collocAN}) = VIF(\texttt{collocNA}) = 2.26$ [26]. La qualité de prédiction de ce modèle est nettement supérieure à celle du Modèle Nul. Tout d'abord, l'exactitude calculée par validation croisée 100 passes est $\mu = 0.873$ ($\sigma = 0.009$). La matrice de confusion correspondante est présentée dans la table 4.23. On observe que ce modèle est capable de prédire plus de 60% des cas d'antéposition, ce qui prouve qu'il modélise le comportement de la variable `position` de façon beaucoup plus satisfaisante que le Modèle Nul.

De plus, l'autre mesure de la qualité du modèle, $AUC = 0.922$, indique que ce dernier permet réellement de faire une prédiction sur le comportement de la variable à prédire. Comme le suggéraient les proportions observées dans la partie précédente, ces deux variables permettent de prédire très largement la position de l'adjectif. L'item nominal avec lequel est combiné l'adjectif, apporte donc une part importante d'informations quant au choix de la position.

26. On a $VIF(\texttt{collocAN}) = VIF(\texttt{collocNA})$, car le modèle ne contient que deux variables.

		Prédits		%
		position=1	position=0	correct
Observés	position=1	2345	1464	61.6%
	position=0	311	9813	96.9%
Exactitude				$\mu = 0.873$ ($\sigma = 0.027$)

TABLE 4.23.: Matrice de confusion du Modèle Collocation

4.3.4. Modèle Lexical

Le Modèle Lexical comprend l'ensemble des informations portées par l'item lexical : sa classe sémantique (`natio`, `intens`, `eval`, `indef`), sa nature morphologique (`morpho` [27]), sa longueur (`longAbs`) et sa fréquence (`freq`) [28]. Le modèle est présenté dans la table 4.24.

	Estimation	Erreur-type	valeur z	Pr(>|z|)	
(Intercept)	-2.56317	0.07185	-35.675	< 2e-16	***
morpho=part	-0.49397	0.17492	-2.824	0.00474	**
morpho=priv	2.75280	0.22604	12.178	< 2e-16	***
morpho=compos	0.10467	0.36451	0.287	0.77400	
morpho=non_constr	1.19959	0.08184	14.657	< 2e-16	***
natio=1	-5.66983	1.00158	-5.661	1.51e-08	***
intens=1	1.50295	0.08867	16.950	< 2e-16	***
eval=1	2.46220	0.14990	16.426	< 2e-16	***
indef=1	2.45060	0.18824	13.019	< 2e-16	***
longAbs	-1.64061	0.08512	-19.275	< 2e-16	***
freq	0.67569	0.02078	32.514	< 2e-16	***

(Significativité des effets selon le test de Wald : *** p<0.001, ** p<0.01, * p<0.05)

TABLE 4.24.: Paramètres du Modèle Lexical

On observe que, hormis la variable `morpho`, les variables votant pour la postposition sont `natio` et `longAbs`. Toutes les autres variables favorisent l'antéposition. Les

27. Rappelons que cette variable est codée de la façon suivante :

	part	priv	compos	non_constr
autres	0	0	0	0
part	1	0	0	0
priv	0	1	0	0
compos	0	0	1	0
non_constr	0	0	0	1

28. Aucune variable n'a été éliminée après avoir effectué des comparaisons de modèles basées sur le test du rapport de vraisemblance.

variables binaires n'expriment leurs préférences que dans le cas où elles sont égales à 1. En ce qui concerne les variables relatives à la fréquence et à la longueur, les préférences sont d'autant plus fortes que leur propre valeur est élevée. Ainsi, plus un adjectif est long, plus le vote pour la postposition est fort ; et plus la fréquence de l'adjectif est importante, plus le vote pour l'antéposition est fort. En ce qui concerne la variable `morpho`, les interprétations ne sont pas tout à fait similaires à celles établies dans la section 4.2.3. En effet, on observe que les participes votent pour la postposition alors que les composés ont une préférence très proche de celle des autres adjectifs construits (valeur `autres` de la variable). De plus, les adjectifs dérivés en *in-* favorisent fortement l'antéposition, tandis que les non-construits ont une préférence moins marquée pour cette position. Ces différences de préférences sont notamment dues à la prise en compte de la longueur. Par exemple, en raison de leur complexité morphologique, les dérivés en *in-* sont longs (3.9 syllabes en moyenne) et, de ce fait, tendent à préférer la postposition. Une fois prise en compte cet effet de longueur, comme c'est le cas dans le Modèle Lexical, on observe que la présence du préfixe *in-* favorise fortement l'antéposition. Les propriétés lexicales des adjectifs peuvent donc avoir des préférences contradictoires pour un même adjectif. Ainsi, en prenant en compte un ensemble de variables relatif aux items adjectivaux dans un modèle unifié, on observe que les adjectifs munis du préfixe privatif présentent une préférence pour l'antéposition, alors que la simple observation des proportions ne permettait pas de le déterminer. Ce résultat est en adéquation avec l'hypothèse formulée dans le chapitre 3, à partir de la littérature sur le phénomène de l'alternance de position de l'adjectif.

La mutlicolinéarité de ce modèle n'est pas très élevée, avec $\kappa = 7.94$ et les facteurs d'inflation de la variance compris entre 1 et 4. Il semble que le κ un peu plus élevé de ce modèle s'explique notamment par la corrélation entre longueur et fréquence.

La qualité de prédiction de ce modèle est très bonne, comme pour le Modèle Collocation. Son exactitude moyenne obtenue par validation croisée 100 passes est $\mu = 0.868$ ($\sigma = 0.030$). Les prédictions exactes sont présentées dans la matrice de confusion en 4.25. Grâce à la prise en compte des caractéristiques lexicales, le modèle prédit correctement près de 75% des antépositions observées.

		Prédits		%
		`position=1`	`position=0`	correct
Observés	`position=1`	2837	972	74.5%
	`position=0`	859	9265	91.5%
Exactitude				$\mu = 0.868$ ($\sigma = 0.030$)

Table 4.25.: Matrice de confusion du Modèle Lexical

La valeur de l'aire sous la courbe ROC confirme la qualité de ce modèle : $AUC = 0.931$. La nature de l'item lexical d'un point de vue de la forme (`longAbs`, `morpho`), de la sémantique (`eval`, `indef`, `intens`, `natio`) mais aussi de l'usage (`freq`) permet

de connaître la position de l'adjectif dans une majorité de cas. Cela amène à penser qu'il serait intéressant de prendre en compte l'identité de chaque adjectif. C'est ce que nous allons faire avec le modèle suivant.

4.3.5. Modèle Lexicalisé

Dans cette section, nous utilisons le modèle de régression logistique à effets mixtes. Pour le problème qui nous intéresse, il se définit de la façon suivante :

$$P(\texttt{position} = 1|X, L_i) = \frac{e^{X\beta+L_i}}{1+e^{X\beta+L_i}} \tag{4.2}$$

où X renvoie aux variables prédictrices constituant les effets fixes et L_i aux effets aléatoires.

Le Modèle Lexicalisé est un modèle à effets mixtes qui contient un effet aléatoire et aucun effet fixe. L'effet aléatoire est le lemme adjectival. Nous le notons L_i, où i représente le lemme considéré. En utilisant les effets aléatoires, on prend en compte le groupement des données autour de chaque lemme adjectival. Plus exactement, chaque lemme se voit associer un coefficient propre qui capte son comportement spécifique. Le Modèle Lexicalisé est présenté dans la table 4.26. Son intercept général est de -2.9, ce qui signifie que tout adjectif ayant un intercept aléatoire inférieur à +2.9 sera prédit postposé.

Effets aléatoires :

Groupes	Nom	Variance	Ecart-type
lem_adj	(Intercept)	4.934	2.2213

Nombre d'obs. : 13933 ; groupes : lem_adj, 1750

Effets fixes :

	Estimation	Erreur-type	valeur z	Pr(>|z|)	
(Intercept)	-2.9059	0.0893	-32.54	<2e-16	***

Table 4.26.: Paramètres du Modèle Lexicalisé

La qualité du Modèle Lexicalisé est excellente : son exactitude moyenne est $\mu = 0.922$ ($\sigma = 0.010$) et l'aire sous la courbe ROC après validation croisée 10 passes est de $AUC = 0.974$ ($\sigma = 0.004$). Le détail des prédictions avec seuil de décision à $P(\texttt{position} = 1|L_i) = 0.5$ est présenté dans la matrice de confusion de la table 4.27.

Si l'on compare les prédictions du Modèle Lexicalisé avec celles du Modèle Lexical, on constate que le Modèle Lexicalisé améliore nettement la prédiction de l'antéposition qui atteint plus de 87%. Il améliore également la prédiction de la postposition en passant à plus de 94%. La valeur AUC suggère que la classification proposée par le Modèle Lexicalisé est très satisfaisante. Dans le Modèle Lexical, les caractéristiques lexicales sont prises en compte en tant qu'effets fixes, alors que

		Prédits		%
		`position`=1	`position`=0	correct
Observés	`position`=1	3324	485	87.3%
	`position`=0	598	9526	94.1%
Exactitude				$\mu = 0.922$ ($\sigma = 0.010$)

Table 4.27.: Matrice de confusion du Modèle Lexicalisé

dans le Modèle Lexicalisé, ces caractéristiques sont introduites comme un effet aléatoire qui affecte chaque lemme adjectival. Dans le premier cas, les aspects lexicaux sont envisagés comme des contraintes générales agissant sur la catégorie grammaticale adjectif, tandis que dans le deuxième cas, on considère qu'il existe un phénomène d'alternance de l'adjectif, mais que ce phénomène doit être envisagé de façon spécifique pour chaque item adjectival, chacun ayant un comportement propre. Le saut qualitatif que permet le passage du Modèle Lexical au Modèle Lexicalisé peut s'expliquer par le fait que le Modèle Lexicalisé tient compte de chaque lemme et de ce fait, a une couverture plus large du phénomène. Pour prendre en compte le comportement spécifique de plus de lemmes dans le Modèle Lexical, il faudrait intégrer chaque lemme à une classe lexicale.

En dépit de sa qualité, le Modèle Lexicalisé pose problème pour la modélisation. Rappelons qu'une part importante des lemmes adjectivaux de notre table de données n'apparaît que dans une seule position (cf. Table 4.2). Cela signifie que, pour ces données, l'identité du lemme permet de connaître la position de l'adjectif de façon catégorique. Or, la présence d'éléments déterminant catégoriquement la valeur de la variable `position` pose des problèmes de convergence pour l'estimation des paramètres du modèle à effets mixtes, en particulier pour l'estimation des intercepts aléatoires associés aux adjectifs qui n'alternent pas dans les données [29]. Pour obtenir une estimation satisfaisante des intercepts aléatoires, il est nécessaire d'écarter les données pour lesquelles aucune alternance n'est observée. Nous avons donc réduit la table de données aux adjectifs présentant une alternance, c'est-à-dire apparaissant au moins une fois dans chaque position. Cette sous-table contient 4994 occurrences d'adjectifs représentant 171 lemmes. Parmi ces adjectifs, 67% sont antéposés et 33% postposés. Nous construisons un Modèle Lexicalisé Alt sur cette sous-table. Comme dans le modèle précédent, le Modèle Lexicalisé Alt ne contient aucun effet fixe. Seul le lemme adjectival est introduit comme effet aléatoire. Le modèle est présenté dans la table 4.28. L'intercept général, proche de 0, indique que seul l'intercept aléatoire

29. Les problèmes de convergence ont été constatés en estimant les paramètres du modèle avec `lmer` sous R. Afin de nous assurer que la non-convergence n'était pas liée à l'algorithme utilisé dans `lmer`, nous avons effectué une simulation de Monte-Carlo à l'aide de JAGS (http://mcmc-jags.sourceforge.net/). Après avoir constaté que le problème de convergence persistait, nous avons conclu que cela était dû à la distribution de nos données. Nous remercions Benoît Crabbé pour son aide à la réalisation de ces manipulations.

associé à chaque lemme permet de déterminer la position de l'adjectif dans ce modèle.

```
Effets aléatoires :
  Groupes  Nom          Variance  Ecart-type
  lem_adj  (Intercept)  2.5391    1.5934
Nombre d'obs. : 4994 ; groupes : lem_adj, 171

Effets fixes :
               Estimation  Erreur-type  valeur z  Pr(>|z|)
  (Intercept)  0.03315     0.13739      0.241     0.81
```

TABLE 4.28.: Paramètres du Modèle Lexicalisé Alt

Les intercepts aléatoires de quelques lemmes relativement fréquents sont présentés dans le graphique de la figure 4.6 et la totalité de ces intercepts aléatoires sont présentés dans la section E.1 de l'annexe E. On observe par exemple que *grand*, *mauvais* et *nouveau* ont des intercepts positifs, ce qui marque leur préférence pour l'antéposition. Les adjectifs *britannique*, *total* et *majeur* ont un intercept négatif qui reflète leur préférence pour la postposition. Les adjectifs *différent* et *important* ont des intercepts proches de 0, ce qui signifie que leur préférence pour une position ou pour l'autre est faible. Ces intercepts aléatoires permettent de comparer, à l'intérieur du modèle, les préférences de chaque lemme. Ainsi, *grand* a une préférence plus forte que *nouveau* pour l'antéposition, et *important* n'a qu'une très légère préférence pour la postposition, comparé aux adjectifs *total* et *moyen*.

Le Modèle Lexicalisé Alt présente une exactitude moyenne de $\mu = 0.798$ ($\sigma = 0.020$) et l'aire sous la courbe ROC après validation croisée 10 passes est $AUC = 0.875$ ($\sigma = 0.016$). Ce modèle prend en compte les aspects lexicaux intervenant dans le choix de la position des adjectifs alternant effectivement dans nos données. L'enjeu est alors de voir si la prise en compte de variables prédictrices mentionnées précédemment permet d'améliorer la modélisation du comportement de la variable `position`.

4.3.6. Modèle Global

Le Modèle Global est construit sur la sous-table de données contenant les adjectifs alternant. Ce modèle combine l'effet aléatoire qui résume les variables lexicales avec les variables du Modèle Syntaxe, celles du Modèle Collocation ainsi que les deux variables concernant la liaison : `conLatMS` et `hiatusPost` [30]. Nous formulons l'hypothèse selon laquelle les intercepts aléatoires associés aux lemmes adjectivaux permettent de résumer les caractéristiques lexicales relatives à chaque adjectif. Autrement dit, nous estimons que l'intercept aléatoire associé à chaque lemme représente une synthèse des

30. Ces deux variables n'ont pas fait l'objet d'un modèle à part. En effet, nous avons montré à travers leur description qu'elles ne concernent qu'un nombre limité de données. Leur utilisation dans un modèle à part ne présentait pas d'intérêt particulier.

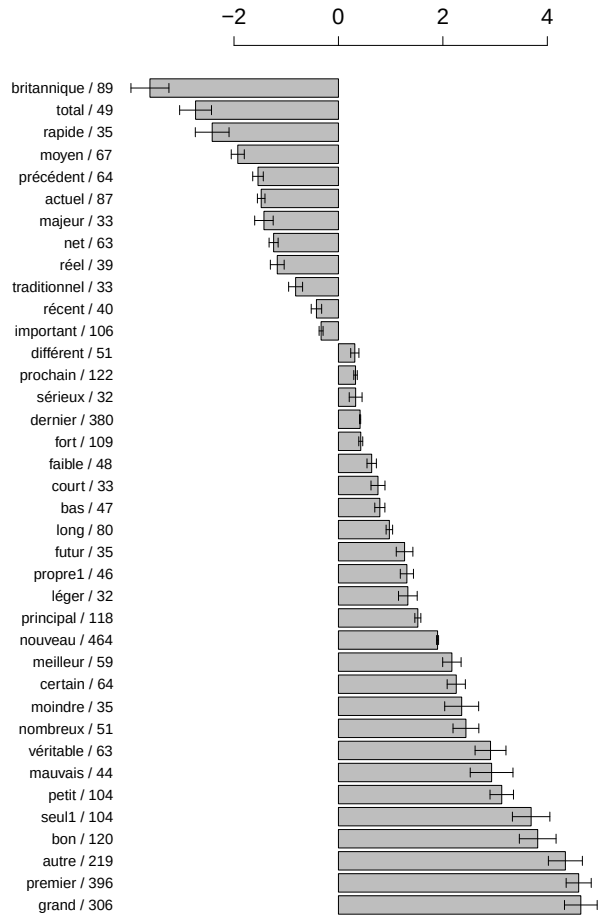

FIGURE 4.6.: Intercepts aléatoires pour un échantillon de lemmes adjectivaux.

différentes contraintes lexicales que sont la longueur, la fréquence, les classes lexicales et la morphologie. D'après cette hypothèse, il n'est pas nécessaire d'introduire les caractéristiques lexicales sous forme d'effets fixes dans le Modèle Global. Cela permet également de réduire la mutlicolinéarité du modèle. En effet, les variables lexicales sont souvent corrélées car elles captent différents aspects d'un même objet et ces derniers sont souvent en lien. Par exemple, un adjectif de nationalité est très souvent un adjectif construit. De même, un adjectif court est souvent un adjectif fréquent.

Le modèle a été compacté par comparaison de modèles sur la base du test de rapport de vraisemblance. Les variables concernant les fonctions syntaxiques (`modifieur`, `ats`, `objet`, `sujet`) ont été éliminées. De même, des variables concernant la configuration du SN (`absDet`, `relative`) ont été écartées car elles ne participaient pas significativement au modèle. Enfin, les deux variables relatives au hiatus et à la liaison (`conLatMS`, `hiatusPost`) ont été éliminées. Cela signifie qu'une fois pris en compte l'item adjectival ainsi que l'item nominal à travers les deux variables de collocation, on n'observe pas d'effet de la fonction du SN ni d'une partie des éléments apparte-

nant au SN. En particulier, on observe que la présence d'une relative ou l'absence de déterminant n'a pas d'effet. On remarque également que la variable `detPoss` qui a été éliminée du Modèle Syntaxe a un effet significatif dans le Modèle Global. Cela signifie que la prise en compte des variables non-syntaxiques a fait émerger l'effet de la présence d'un déterminant possessif. Enfin, en ce qui concerne les variables relatives au hiatus et à la liaison, elles ne sont pas non plus significatives. Étant donné la nature de nos données, ce résultat ne peut être vu que comme un indice du fait que ce type de contraintes n'intervient pas à l'écrit [31].

Le modèle compacté contient dix variables. Il est présenté dans la table 4.33 [32]. Les variables syntaxiques `artDef`, `detDem`, `detPoss`, `sprep`, `adjAnt` et `adjPost` votent pour l'antéposition, tandis que `adv` et `coord` favorisent la postposition. Les deux variables relatives aux collocations votent en fonction de leur propre valeur : plus leur valeur est élevée, plus leur préférence est marquée. Comme attendu, `collocAN` favorise l'antéposition et `collocNA` la postposition. L'intercept général du modèle (-0.29) favorise légèrement la postposition. Les valeurs des intercepts aléatoires associées aux lemmes adjectivaux sont présentées dans la section E.2 de l'annexe E. Ce modèle n'est pas affecté par des problèmes de multicolinéarité, comme en témoigne la valeur très faible de l'indice de conditionnement : $\kappa = 4.27$.

La comparaison entre les proportions présentées dans la partie 4.2 et les variables significatives dans le Modèle Global montre l'intérêt de la modélisation que nous proposons. En effet, on observe par exemple que l'article défini qui semblait favoriser la postposition (76.5%), vote légèrement pour l'antéposition d'après le Modèle Global. Grâce à cette modélisation, on constate également que les variables concernant la fonction du SN que Forsgren (1978) avait mises à jour dans son étude de corpus, ne sont pas significatives lorsque les variables qui expliquent véritablement le choix de la position de l'adjectif sont prises en compte. La machinerie statistique déployée semble donc être un outil de modélisation adéquat pour rendre compte d'un phénomène aussi complexe que celui qui nous occupe.

Le modèle a de bonnes capacités à prédire la position de l'adjectif : son exactitude au seuil de décision $P(\texttt{position} = 1|X, L_i) = 0.5$ est de $\mu = 0.869$ ($\sigma = 0.015$). De plus, la mesure AUC après validation croisée 10 passes est de 0.935. On observe une nette amélioration des mesures de qualité par rapport au Modèle Lexicalisé Alt, qui porte sur le même sous-ensemble d'adjectifs. Cela indique que les contraintes générales considérées dans le Modèle Global permettent de mieux décrire la position de l'adjectif épithète, une fois les préférences lexicales prises en compte. Le détail des prédictions donné dans la matrice de confusion (table 4.29) montre que le Modèle Global prédit correctement, pour les adjectifs alternant, l'antéposition à près de 93% et la postposition à 80.1%.

Le graphique en 4.7 représente l'ajustement des proportions moyennes de données

31. La conjecture selon laquelle le hiatus et la liaison n'agissent pas à l'écrit, devrait être étayée empiriquement. En effet, on sait que les locuteurs respectent les contraintes prosodiques de la langue quand ils lisent (Fodor, 2002). On peut se demander si ces contraintes sont également actives lorsqu'ils écrivent.

32. En raison de sa taille, cette table se trouve à la fin du chapitre.

		Prédits		%
		`position=1`	`position=0`	correct
Observés	`position=1`	3110	237	92.9%
	`position=0`	328	1319	80.1%
Exactitude				$\mu = 0.869$ ($\sigma = 0.015$)

TABLE 4.29.: Matrice de confusion du Modèle Global

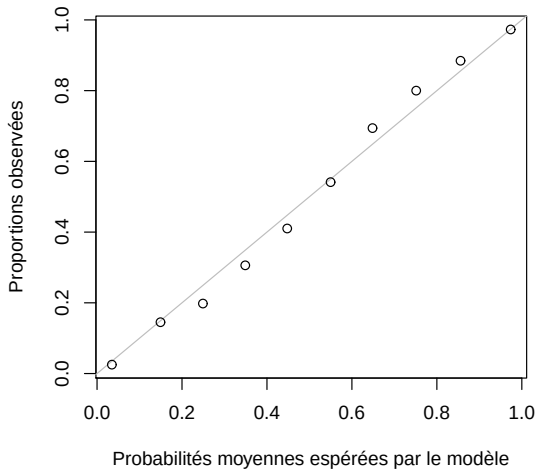

FIGURE 4.7.: Ajustement des données observées groupées en fonction des probabilités prédites moyennes pour le Modèle Global.

observées en fonction des probabilités prédites. La droite représente un ajustement parfait. La répartition des points le long de cette droite montre que le Modèle Global est bien ajusté aux données.

L'ensemble des indicateurs prouve que le Modèle Global permet de modéliser l'alternance de position de l'adjectif avec une bonne qualité de prédiction et un bon ajustement aux données.

4.4. Bilan

Dans cette étude sur corpus de la position de l'adjectif en français, nous avons mis au point un modèle permettant de prédire la position des adjectifs présentant une alternance dans nos données à près de 87%.

Nous illustrons la qualité de la prédiction en présentant deux exemples concrets associés à leur probabilité. Le premier concerne l'adjectif *fort* qui a un intercept

aléatoire de $-0,097$. Étant donné l'intercept général du Modèle Global (-0.29), cet adjectif présente une très légère préférence pour la postposition, si toutes les variables prédictrices sont égales à 0. Pour l'exemple présenté en (10), le Modèle Global attribue la probabilité $P(\texttt{position} = 1|X, L_{fort}) = 0.294$.

(10) *Mais cela suppose un changement dans les mentalités, la volonté réelle d'acquérir ce que les spécialistes appellent une culture de sûreté et celle, tout aussi importante, de mettre en place une autorité de sûreté* ***forte*** *et indépendante.*

La présence du SP *de sûreté* favorise l'antéposition, mais la coordination de l'adjectif vote plus fort pour la postposition, ce qui explique la probabilité obtenue. Le deuxième exemple concerne l'adjectif *proche* qui a un intercept aléatoire de -0.621, ce qui signifie qu'il a une préférence pour la postposition. Dans l'exemple (11), le SN est introduit par un possessif, ce qui favorise l'antéposition d'après le Modèle Global. De plus, la séquence ordonnée *proche collaborateur* a une valeur élevée pour `collocAN` ($= 11.14$). Malgré la préférence de l'adjectif pour la postposition, le modèle assigne donc la probabilité $P(\texttt{position} = 1|X, L_{proche}) = 0.942$ à cet exemple.

(11) *Il écrivait en début de semaine : "Au vu des promesses faites aux salariés de la Cinq et à leur famille, il y a un an, et du non-respect de ces engagements moraux, la CFTC Bourse estime qu'une entreprise qui a si peu de parole vis-à-vis de ses* ***proches*** *collaborateurs n'a aucune raison d'en avoir plus vis-à-vis de ses actionnaires…"*

À travers la comparaison de modèles, nous avons analysé des contraintes relevant de différents niveaux d'organisation. Nous avons d'abord montré l'importance de l'item adjectival dans ce phénomène. La tendance observée dans le Modèle Lexical est que converge vers chaque position un faisceau de caractéristiques lexicales formelles. Ainsi, les adjectifs antéposés ont tendance à être courts, fréquents et non-construits, tandis que les adjectifs postposés ont tendance à être longs, moins fréquents et construits. Cette convergence générale de caractéristiques lexicales peut se représenter sous la forme d'un schéma, comme dans la table 4.30. Le Modèle Lexical a également permis de montrer que les caractéristiques lexicales formelles engendrent des préférences contradictoires pour un même lemme : les adjectifs munis du préfixe privatif *in-* ont une préférence pour l'antéposition, alors que leur longueur, plus élevée que la moyenne, favorise la postposition.

Antéposition	NOM	**Postposition**
court		long
fréquent		rare
simple		construit

TABLE 4.30.: Convergence de faisceaux de caractéristiques lexicales selon les positions

À ces caractéristiques lexicales constitutives de la forme de l'item, s'ajoutent des caractéristiques relevant de la sémantique. Le Modèle Lexical ne contient que quatre classes lexicales mais les tendances sont très claires. Les adjectifs intensionnels, évaluatifs et indéfinis préfèrent l'antéposition, alors que les adjectifs de nationalité préfèrent la postposition. Cependant, le classement systématique des adjectifs est problématique, car la définition des contours d'une classe sémantique ne va pas de soi. C'est en partie pour cette raison que le Modèle Lexicalisé apparaît plus approprié : en approximant les préférences individuelles des adjectifs, il permet de prendre en compte leur dimension lexicale sans opérer de catégorisation systématique. Il est important de noter que nous avons écarté de notre étude les adjectifs apparaissant avec un dépendant, car la présence d'un constituant postadjectival impose catégoriquement la postposition. Cependant, la possibilité d'introduire un dépendant est une propriété relevant de l'adjectif. Par exemple, à la différence des adjectifs *strict* ou *exceptionnel*, les adjectifs *susceptible* et *nécessaire* peuvent sous-catégoriser un complément, comme dans *susceptible de réussir* ou *nécessaire au bonheur*. Ainsi, la présence d'un dépendant postadjectival relève en partie de la dimension lexicale [33], ce qui renforce l'importance de l'item adjectival dans le choix de la position [34].

D'après la modélisation proposée dans le Modèle Global, une fois considérées les particularités de chaque item lexical, trois niveaux sont significatifs pour rendre compte de la position de l'adjectif : 1) un niveau combinatoire local où l'on considère l'association de deux items lexicaux ; 2) le niveau syntaxique concernant la configuration du SAdj ; 3) un niveau syntaxique plus large englobant l'intégralité du SN.

Le premier niveau a une très grande importance. Le Modèle Collocation comme le Modèle Global montrent que pour connaître la position de l'adjectif, il faut prendre en compte l'item nominal. L'idée est que l'adjectif a une préférence individuelle générale plus ou moins marquée et que la combinaison de l'adjectif avec un nom spécifique peut aller à l'encontre de cette préférence.

Ce résultat suggère que la position de l'adjectif ne doit pas être envisagée seulement au niveau lexical et au niveau syntaxique, mais également à un niveau intermédiaire de combinaison des items. L'utilisation de mesure d'association reposant sur la fré-

33. Notons que la présence d'un constituant en position postadjectivale ne se limite pas aux adjectifs sous-catégorisant un complément. Par exemple, la préposition *comme* peut introduire un constituant postadjectival avec un grand nombre d'adjectif.

34. On pourrait émettre l'hypothèse selon laquelle la possibilité pour un adjectif de sous-catégoriser un complément favorise la postposition. Étant donné que la présence d'un complément impose la postposition, il est envisageable que, sous l'influence de leur position avec complément réalisé, ces items adjectivaux aient tendance à être postposés dans les cas où leur complément n'est pas réalisé. En nous appuyant sur les cadres de sous-catégorisation des adjectifs du Lexique des Formes Fléchies du Français (Le*fff*) (Sagot, 2010), acquis à partir du FTB et corrigés à la main, nous avons créé deux groupes d'adjectifs dans nos données : (1) ceux présentant la possibilité de sous-catégoriser un complément et (2) ceux ne présentant pas cette possibilité. D'après l'hypothèse formulée précédemment, on s'attend à ce que les adjectifs du groupe (1) présentent une proportion de postposition significativement supérieure à celle des adjectifs du groupe (2). Or, parmi les 148 adjectifs du groupe (1) qui représentent 2290 occurrences, on observe 977 cas de postposition, soit 42.7%. La possibilité pour un adjectif de sous-catégoriser un complément semble plutôt favoriser l'antéposition, ce qui va à l'encontre de l'hypothèse de départ.

quence montre que ce niveau intermédiaire est fortement influencé par la fréquence d'apparition du nom et de l'adjectif dans un ordre spécifique. Ainsi, une séquence Nom - Adjectif très fréquemment produite dans un ordre donné va avoir tendance à être reproduite dans cet ordre-là, même si la préférence générale de l'adjectif est différente. La préférence associée à la séquence Nom - Adjectif devient plus importante que la préférence de l'adjectif à mesure que la séquence ordonnée est produite. Cela est particulièrement vrai pour les adjectifs antéposés. Pour la plupart des séquences ordonnées Adjectif - Nom ayant un score d'association élevé (cf. figure 4.4), l'antéposition semble relever d'une convention d'usage, d'une façon de dire les choses. Cela s'observe aussi en postposition, avec une séquence comme *futur proche* qui présente une valeur élevée pour la variable `collocNA`. La séquence inversée *proche futur* est également possible mais présente un score `collocAN` bien inférieur à celui de `collocNA`.

Le deuxième niveau concerne la complexité du SAdj. Lorsque le SAdj est complexe, c'est-à-dire qu'il n'est pas seulement composé d'un adjectif, il a tendance à être postposé. La complexité du SAdj est corrélée à sa longueur. Le placement des adjectifs suit donc une tendance générale du français et des langues à ordre VO (Hawkins, 1994) : lorsque le constituant est long et complexe, il a tendance à apparaître en dernier. Il est intéressant de constater que dans un phénomène aussi marqué par le niveau lexical, la tendance à repousser les éléments complexes en dernier est globalement respectée. Cela suggère que cette tendance est très forte en français.

Le troisième niveau a une importance moindre, mais il reste significatif une fois que les deux autres niveaux sont pris en compte. Il concerne d'une part le déterminant introduisant le SN, et d'autre part, les autres dépendants du nom présents dans le SN. Les déterminants définis, possessifs et démonstratifs favorisent l'antéposition. Le rôle du démonstratif peut être mis en lien avec la contrainte relative à la reprise anaphorique (cf. section 3.7). En effet, le démonstratif est un introducteur privilégié de reprise anaphorique. Ainsi, dans l'exemple (12), l'antéposition de l'adjectif *volumineux*, en partie captée dans le modèle par la contrainte relative au démonstratif, renvoie à un cas de reprise anaphorique.

(12) *Le Conseil supérieur de l'audiovisuel […] a reçu, jeudi 31 décembre, un dossier complémentaire du projet de chaîne éducative Eurêka, en vue de l'attribution de la partie diurne du cinquième canal national de télévision. […] Après examen de* ***ce volumineux dossier****, le CSA pourrait statuer en janvier.*

La tendance à chercher l'équilibre dans le SN en rejetant l'adjectif en antéposition lorsque d'autres éléments sont présents en postposition (`adjPost`, `sprep`) est confirmée par le Modèle Global. Cependant, on observe que la présence d'un autre adjectif antéposé (`adjAnt`) ne favorise pas la postposition comme attendu, mais plutôt l'antéposition. On peut supposer que ce type de contrainte est pertinent dans un corpus journalistique, car les locuteurs ont le temps de planifier leurs phrases et de les modifier pour obtenir des SN équilibrés. Il est possible que ces contraintes soient moins pertinentes pour une modélisation de production orale spontanée.

L'intérêt majeur de la présente contribution au problème de la position de l'adjectif en français est que le Modèle Global permet de combiner les préférences lexicales aux préférences liées aux items nominaux et aux contraintes syntaxiques. On observe notamment que les contraintes syntaxiques ne présentent pas une bonne qualité de prédiction prises isolément (Modèle Syntaxe) ; par contre, combinées aux contraintes plus massives, elles se révèlent significatives.

Ce chapitre a été l'occasion de décrire et de modéliser le phénomène d'alternance de position de l'adjectif à partir de données de corpus. Cependant, ce travail pose un problème de généralisation : peut-on considérer que les préférences observées en corpus correspondent à des préférences chez les locuteurs du français ? Pour tenter de donner un élément de réponse à cette question, nous avons mis en place un questionnaire inspiré de celui utilisé par Bresnan (2007b), qui est décrit dans la section 2.3.2 du chapitre 2.

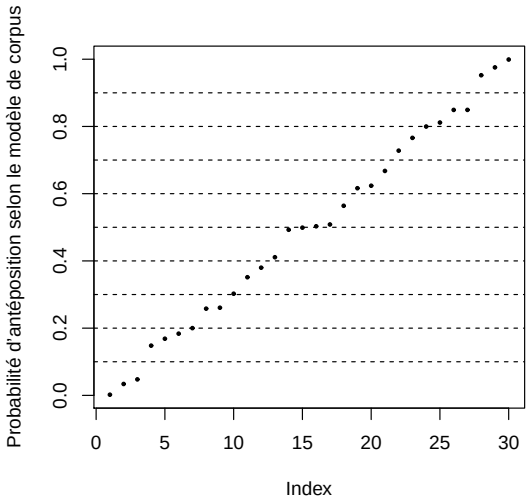

FIGURE 4.8.: Probabilité des 30 phrases utilisées dans le questionnaire.

Préférences sur des paires de phrases et corrélation avec le Modèle Global

L'objectif du questionnaire est de confronter les probabilités prédites par le modèle aux préférences des locuteurs et de voir s'il existe une corrélation entre les deux. L'hypothèse de départ est que, pour un grand nombre de sujets, la fréquence du choix de l'antéposition doit être en correspondance avec la probabilité d'antéposition estimée par le modèle de corpus. Ainsi, pour une phrase ayant une probabilité très faible d'antéposition, très peu de locuteurs devraient choisir la version avec antéposition. Et inversement, pour une phrase présentant une probabilité élevée, la grande majorité des sujets interrogés devrait opter pour l'antéposition. Pour tester cette hypothèse, nous avons mis en place un questionnaire où les sujets devaient choisir l'ordre qu'ils

préféraient pour la séquence Nom-Adjectif en contexte [35].

Le questionnaire se compose de 30 phrases extraites du FTB. Elles ont été sélectionnées en fonction de la probabilité qui leur a été attribuée par le Modèle Global. Nous avons choisi trois phrases pour chaque intervalle de 0.1 point, comme cela est montré dans la figure 4.8.

Pour chaque phrase, les deux options, adjectif antéposé ou postposé, sont présentées au sujet qui doit simplement cocher la case correspondant à sa version préférée. Les phrases sont présentées sous la forme de paires, pour lesquelles seule la position de l'adjectif diffère. Le syntagme nominal est mis en gras et en couleur dans les deux versions de la phrase, afin que le sujet puisse repérer la différence entre les deux versions et puisse ainsi facilement choisir celle qui a sa préférence. Un exemple de paire de phrases à juger est présenté dans la figure 4.9.

☐ *A Bruxelles, où il fut commissaire de 1985 à 1989, chargé d'abord des affaires sociales et de l'éducation, puis responsable de la politique de concurrence, Peter Sutherland semble faire l'unanimité. Aussi est-ce avec* **une satisfaction grande** *que sa désignation à la tête du GATT y a été accueillie, comme si cet Irlandais qui a réussi était paré de toutes les qualités pour s'acquitter avec efficacité et équité de cette mission difficile.*

☐ *A Bruxelles, où il fut commissaire de 1985 à 1989, chargé d'abord des affaires sociales et de l'éducation, puis responsable de la politique de concurrence, Peter Sutherland semble faire l'unanimité. Aussi est-ce avec* **une grande satisfaction** *que sa désignation à la tête du GATT y a été accueillie, comme si cet Irlandais qui a réussi était paré de toutes les qualités pour s'acquitter avec efficacité et équité de cette mission difficile.*

FIGURE 4.9.: Exemple de paire de phrases à juger dans le questionnaire d'élicitation de préférences.

Pour chaque phrase, le sujet doit sélectionner la version qui lui semble la plus appropriée. Dans chaque questionnaire, l'ordre dans lequel sont présentées les phrases, ainsi que l'ordre des deux options, sont déterminés aléatoirement. Les consignes de cette étude corrélationnelle sont reproduites en annexe A, ainsi que les 30 phrases à juger.

35. Dans un premier temps, nous avons construit un questionnaire sur le modèle de celui de Bresnan (2007b), en demandant aux sujets de répartir 100 points sur les deux ordres possibles (antéposition et postposition) en fonction de leur préférence, selon la méthode détaillée dans la section 2.3.2 du chapitre 2. Cependant, nous avons observé une variabilité extrême des données préliminaires recueillies : chacune des trente phrases soumises au jugement de 20 personnes s'est vu attribuer, au moins une fois, un score de 0 et un score de 100. Nous en avons conclu que la tâche qui consiste à répartir des points sur deux options était mal comprise ou, en tout cas, interprétée de façon différente selon les sujets. Nous avons donc opté pour une schéma expérimental plus simple, selon lequel les sujets devaient choisir l'ordre qu'ils préféraient parmi les deux possibles.

L'étude corrélationnelle s'est déroulée via un site internet auquel les sujets ont accédé par un lien diffusé sur des réseaux sociaux et des listes de diffusion scientifique. Nous avons recueilli les réponses de 141 locuteurs natifs du français [36]. Nous avons constaté qu'un des syntagmes nominaux que nous avions soumis au jugement des locuteurs était ambigu, dans la mesure où les deux mots lexicaux pouvaient être analysés comme nom ou comme adjectif. Ce syntagme est présenté en (13) avec l'analyse que nous attendions dans le questionnaire.

(13) *les britanniques*$_{ADJ}$ *responsables*$_{NOM}$ / *les responsables*$_{NOM}$ *britanniques*$_{ADJ}$

Il semble que la façon dont nous avons présenté les données, à savoir la mise en relief du SN, a conduit un certain nombre de sujets à produire l'analyse non attendue. Cette phrase a été écartée de nos analyses. Les résultats que nous donnons sont donc valables pour 29 phrases.

Nous observons l'existence d'une corrélation entre la proportion de choix pour la version antéposée et la probabilité d'antéposition du Modèle Global. Cette corrélation, estimée numériquement avec le coefficient de Pearson, est significative : 0.74 ($p= 3.7 \times 10^{-06}$). Dans la figure 4.10, le graphique représente, pour chaque phrase, la probabilité d'antéposition en fonction de la proportion d'antéposition dans les choix des sujets. La droite de régression montre que la corrélation a une bonne allure.

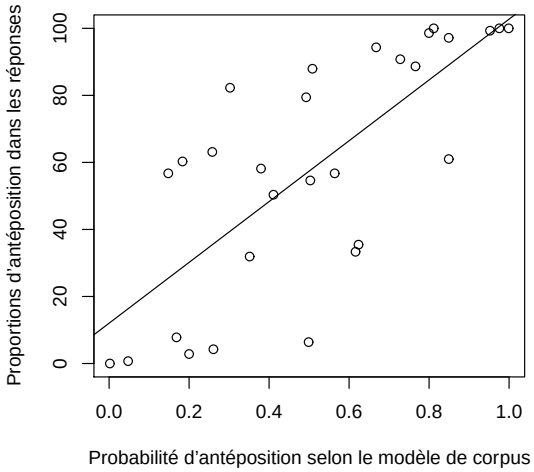

FIGURE 4.10.: Proportions d'antéposition pour les 29 phrases testées en fonction des probabilités d'antéposition estimées par le Modèle Global.

On observe que la droite est décalée au niveau de l'ordonnée à l'origine, par rapport à ce qui serait attendu idéalement. Cela signifie que les sujets ont eu tendance à

36. Cent quatre-vingt-dix personnes ont répondu au questionnaire. Nous avons écarté les personnes ayant une langue maternelle autre que le français et celles qui n'avaient pas rempli entièrement le questionnaire.

préférer l'antéposition plus que ce qui était théoriquement attendu d'après le modèle. Cela est confirmé par la proportion de choix pour l'antéposition par l'ensemble des sujets sur la totalité des phrases : 58.7%. On peut se demander si le style des phrases proposées n'a pas eu une influence sur cette sur-représentation de l'antéposition. En effet, les phrases du journal *Le Monde* présentent un style soutenu qui pourrait avoir conduit les sujets à choisir plus fréquemment l'antéposition.

De plus, certaines des phrases présentent des proportions d'antéposition relativement éloignées des probabilités calculées en corpus. Nous détaillons les trois exemples de ce type.

Dans la phrase (14), l'ordre Adjectif-Nom se voit attribuer une probabilité de 0.15 par le Modèle Global, alors que les sujets de l'étude corrélationnelle ont choisi cet ordre dans 56.7% des cas. Dans le Modèle Global la postposition est très fortement privilégiée par la présence de la coordination. Or, il apparaît qu'en termes de préférence, la coordination ne provoque pas un choix massif pour la postposition. Il semble donc que la coordination, qui a un effet marqué dans les données du FTB, ait un effet bien moindre dans le cas de préférences élicitées.

(14) a. *Encore faudrait-il que, pour faire passer la pilule des réformes nécessaires - et que beaucoup d'Italiens, en dépit de leur enthousiasme, risquent, une fois au pied du mur, de trouver plus amère que prévu, - qu'**un fort et surtout crédible gouvernement** se constitue.*

b. *Encore faudrait-il que, pour faire passer la pilule des réformes nécessaires - et que beaucoup d'Italiens, en dépit de leur enthousiasme, risquent, une fois au pied du mur, de trouver plus amère que prévu, - qu'**un gouvernement fort et surtout crédible** se constitue.*

En ce qui concerne la phrase (15), la probabilité d'antéposition est de 0.18. Les sujets de l'étude corrélationnelle ont opté pour cette position dans 60% des cas. La raison de cette préférence pour la position antéposée ne nous paraît pas claire. Il est possible que les guillemets, qui font partie de la typographie originale du texte, aient influencé la décision des locuteurs.

(15) a. *Cette solution vise à désamorcer les craintes du Congrès quant à un transfert **de "sensibles technologies"** dans des mains étrangères, en l'occurrence françaises.*

b. *Cette solution vise à désamorcer les craintes du Congrès quant à un transfert **de "technologies sensibles"** dans des mains étrangères, en l'occurrence françaises.*

Enfin, pour la phrase (16), les sujets ont choisi à plus de 93.6% la postposition, alors que la probabilité pour l'antéposition estimée par le modèle est de 0.49. Les données de corpus tendent à indiquer que l'adjectif *difficile* peut apparaître dans les deux positions de façon équivalente, alors que les sujets ont une préférence très marquée pour la postposition.

(16) a. *Il se demande où il va loger sa famille de trois enfants s'il a la malchance de vivre en région parisienne, combien d'heures de trajet il subira chaque jour pour aller travailler et* ***combien de fins de mois difficiles*** *il devra affronter.*

b. *Il se demande où il va loger sa famille de trois enfants s'il a la malchance de vivre en région parisienne, combien d'heures de trajet il subira chaque jour pour aller travailler et* ***combien de difficiles fins de mois*** *il devra affronter.*

La corrélation constatée et l'allure de la droite de régression montrent qu'il existe une correspondance entre les probabilités évaluées en corpus et les préférences que les locuteurs peuvent avoir dans une tâche métalinguistique. Cela constitue un argument pour affirmer que les observations faites sur la base de fréquences en corpus ont une correspondance avec une forme de savoir langagier. Cependant, comme nous l'avons illustré, les données expérimentales sont, dans certains cas, éloignées des donnés attendues d'après le Modèle Global. Le protocole expérimental que nous avons mis en place pourrait être amélioré. Premièrement, nos observations relatives à la préférence des locuteurs s'appuient sur la fréquence des choix pour l'antéposition. Il est peut-être nécessaire d'augmenter le nombre de participants pour accroître la qualité des résultats. Deuxièmement, le questionnaire a été rempli en ligne. Nous n'avions donc aucun contrôle sur les conditions d'expérimentation. Ce type d'expérience nécessite peut-être des conditions expérimentales plus homogènes. Enfin, les phrases qui ont été soumises au jugement sont des phrases extraites du journal *Le Monde* traitant de sujets économiques et politiques. Rappelons que dans le cas du travail de Bresnan (2007b), les données jugées étaient des données d'oral. Il serait intéressant de conduire la même étude corrélationnelle avec des phrases d'oral ou d'écrit plus relâché.

Antéposition et préférences lexicales dans des corpus oraux et littéraires

Le travail présenté dans ce chapitre repose sur l'analyse de données issues de l'écrit journalistique. Afin d'avoir une idée de la variation de la position de l'adjectif selon le genre de corpus, nous avons relevé le pourcentage d'antéposition pour les 171 adjectifs alternant de nos données dans un corpus littéraire ainsi que dans deux corpus d'oral. L'hypothèse de départ est que l'alternance est moins forte à l'oral que dans le genre journalistique, mais qu'elle est plus forte dans le genre littéraire.

En ce qui concerne le genre littéraire, nous avons repris les comptes d'occurrences proposés par Wilmet (1980). Rappelons que ce corpus est composé des adjectifs épithètes extraits de 80 oeuvres littéraires du XX^{e} siècle. Il est important de noter que dans ce corpus, les dépendants postadjectivaux n'ont pas été écartés. Dans la mesure où la proportion d'adjectifs ayant un dépendant représente environ 3.5% des données que nous avons extraites du FTB, on peut estimer que la proportion de dépendants postadjectivaux est du même ordre dans le corpus de Wilmet (1980). De

plus, les adjectifs homonymes, tels que *ancien*, *propre*, *pur*, *seul* et *simple*, ne sont pas désambiguïsés. Enfin, cinq adjectifs ne sont pas attestés : *mirobolant*, *influent*, *prétendu*, *aléatoire* et *moyen*. Dans le corpus de Wilmet, les 171 adjectifs représentent 9275 occurrences, parmi lesquelles 6197 (66,8%) sont antéposées.

	Nombre total d'occ.	Antéposés	
FTB	4994	3347	67.0%
ESTER	2543	1680	66.0%
CORAL-ROM	1342	906	67.5%
Wilmet	9275	6197	66.8%

TABLE 4.31.: Proportions d'antéposition dans les corpus FTB, ESTER, CORAL-ROM ainsi que dans les données de Wilmet (1980) pour les 171 adjectifs alternant.

Pour ce qui est du genre oral, nous avons extrait les 171 adjectifs des corpus ESTER et CORAL-ROM, qui sont présentés dans la section 2.1.3 du chapitre 2. Le premier est un corpus radiophonique transcrit et le deuxième contient des conversations et des monologues en contexte formel et informel. Nous avons relevé un échantillon de 100 occurrences par adjectif et par corpus. Dans les cas où l'adjectif présentait moins de 100 occurrences, nous avons extrait la totalité de ses occurrences. Comme pour les données littéraires de Wilmet, les deux sens des adjectifs homonymes n'ont pas été différenciés et une partie des adjectifs n'est pas attestée : 25 non-attestés dans ESTER et 40 dans CORAL-ROM. Nous avons ainsi recueilli :

- dans le corpus ESTER, 2543 occurrences d'adjectifs avec une proportion d'antéposition de 66% (1680 occurrences) ;
- dans le corpus CORAL-ROM, 1342 occurrences, dont 906 (67.5%) sont antéposées par rapport au nom.

Les proportions d'antéposition dans les données des trois corpus (Wilmet, ESTER et CORAL-ROM) sont proches de celles rencontrées dans les données extraites du FTB, comme le montre le tableau 4.31.

Afin de donner une idée des préférences lexicales, nous comparons le pourcentage d'occurrences antéposées pour les adjectifs présentant plus de 20 occurrences dans chaque corpus [37]. Le tableau 4.32 permet d'observer que pour les adjectifs présentant une préférence marquée pour l'antéposition (mis en gras dans le tableau), il n'y a pas de tendance opposée selon les corpus : les proportions d'adjectifs antéposés présentent le même ordre de grandeur. En revanche, les proportions relatives aux deux adjectifs ayant plutôt une préférence pour la postposition sont plus disparates. Les adjectifs *important* et *différent* sont surreprésentés en antéposition dans le FTB par rapport aux autres corpus. Ainsi, l'hypothèse selon laquelle l'alternance est plus forte dans le genre littéraire et moins forte à l'oral ne semble pas vérifiée

37. Le nombre d'occurrences antéposées et postposées pour les 171 adjectifs dans les quatre corpus (FTB, ESTER, CORAL-ROM et Wilmet) est présenté dans la section E.3.1 l'annexe E.

	FTB			ESTER		
Adjectif	Total	Antéposés		Total	Antéposés	
grand	306	305	99.7%	99	96	97.0%
autre	219	218	99.5%	100	99	99.0%
bon	120	119	99.2%	100	98	98.0%
petit	104	101	97.1%	99	99	100.0%
seul	112	104	92.9%	77	77	100.0%
certain	64	59	92.2%	53	53	100.0%
meilleur	59	54	91.5%	37	36	97.3%
dernier	380	232	61.1%	100	78	78.0%
nouveau	464	406	87.5%	99	95	96.0%
important	106	45	42.5%	71	11	15.5%
différent	51	30	58.8%	31	14	45.2%
	CORAL-ROM			Wilmet		
Adjectif	Total	Antéposés		Total	Antéposés	
grand	100	95	95.0%	1304	1262	96.8%
autre	100	99	99.0%	68	65	95.6%
bon	45	45	100.0%	479	467	97.5%
petit	100	99	99.0%	1139	1124	98.7%
seul	38	37	97.4%	247	210	85.0%
certain	37	36	97.3%	46	41	89.1%
meilleur	22	20	90.9%	30	28	93.3%
dernier	58	35	60.3%	116	88	75.9%
nouveau	39	31	79.5%	221	141	63.8%
important	44	6	13.6%	21	4	19.0%
différent	39	13	33.3%	39	9	23.1%

TABLE 4.32.: Nombre d'occurrences et pourcentage d'antéposition dans les corpus FTB, ESTER, CORAL-ROM ainsi que dans les données de Wilmet (1980) pour onze adjectifs.

d'après nos premières observations. En effet, la proportion d'adjectifs antéposés est quasi-équivalente dans les quatre corpus et la comparaison des proportions item par item ne montre pas une prédominance de l'antéposition dans le genre littéraire par rapport aux genres journalistique et oral.

Nous sommes partie de l'hypothèse que l'alternance de position est valable pour la classe des adjectifs et qu'elle est guidée par des contraintes préférentielles. Le modèle à effets aléatoires indique que la seule connaissance de l'item adjectival permet de connaître la position dans 92% des cas. L'alternance de position n'est donc pas massive. Cependant, il reste une part d'alternance, guidée par des contraintes préférentielles, dont nous avons rendu compte dans le Modèle Global. Ces contraintes concernent l'item nominal mis en jeu, la structure du SAdj et la configuration du SN. Le questionnaire que nous avons fait remplir montre que les préférences modélisées sur corpus sont corrélées à des tendances observées dans le choix des locuteurs.

Ce travail doit se poursuivre selon deux perspectives. Premièrement, la modélisation sur corpus journalistique que nous avons proposée devrait être étendue à d'autres genres de discours, afin de pouvoir évaluer la généralité du Modèle Global et de comparer les tendances de chaque genre. Les premiers chiffres obtenus à partir de corpus oraux et littéraires semblent indiquer que la variation entre les genres n'est pas particulièrement importante au niveau lexical. Ces observations doivent être confirmées et les contraintes non-lexicales doivent être étudiées pour ces genres.

Deuxièmement, la recherche de corrélation entre les observations en corpus et les préférences des locuteurs doit être approfondie. En effet, il est nécessaire de développer des arguments permettant de renforcer l'idée selon laquelle les résultats obtenus grâce aux méthodes d'analyse de données en corpus sont en correspondance avec une forme de savoir langagier. Dans cet esprit, nous avons présenté un questionnaire reposant sur l'élicitation de préférences. Nous pourrions envisager d'autres tâches, notamment en compréhension, fondées, par exemple, sur les techniques d'auto-présentation segmentée ou *self-paced reading*.

Effets aléatoires :

Groupes	Nom	Variance	Ecart-type
lem_adj	(Intercept)	2.1367	1.4618

Nombre d'obs. : 4994 ; groupes : lem_adj, 171

Effets fixes :

	Estimation	Erreur-type	valeur z	Pr(>|z|)	
(Intercept)	-0.29327	0.15559	-1.885	0.059452	
artDef=1	0.38677	0.11436	3.382	0.000720	***
detDem=1	1.53952	0.28554	5.392	6.98e-08	***
detPoss=1	0.96312	0.25320	3.804	0.000142	***
coord=1	-1.35389	0.28564	-4.740	2.14e-06	***
adjAnt=1	0.55510	0.26261	2.114	0.034531	*
adjPost=1	0.58506	0.16017	3.653	0.000259	***
sprep=1	0.86851	0.11127	7.805	5.94e-15	***
adv=1	-1.70414	0.18981	-8.978	< 2e-16	***
collocNA	-0.44146	0.02207	-20.006	< 2e-16	***
collocAN	0.36458	0.01997	18.256	< 2e-16	***

Corrélation des effets fixes :

	(Int)	def	dem	poss	coo	aAnt	aPos	sp	adv	NA
def=1	-.270									
dem=1	-.131	.209								
poss=1	-.149	.236	.093							
coord=1	-.099	.013	-.016	-.003						
aAnt=1	-.035	.003	.050	.013	.008					
aPost=1	-.196	-.037	.020	.028	.038	.016				
sprep=1	-.293	-.102	.052	.046	.007	.028	.173			
adv=1	-.064	-.075	-.009	-.069	.050	-.028	-.052	-.026		
colNA	-.119	-.050	-.028	-.021	.065	-.077	.072	.028	.126	
colAN	-.165	.083	.001	.034	-.057	.032	.097	.083	-.129	-.508

TABLE 4.33.: Paramètres du Modèle Global

Deuxième partie .

Les compléments postverbaux

Chapitre 5

L'ordre des dépendants du verbe - État de l'art

Sommaire

Dans ce chapitre, nous nous intéressons à l'ordre des compléments du verbe dans le domaine postverbal. En français, l'ordre de deux compléments sous-catégorisés par le verbe est relativement libre. Dans l'exemple qui suit, la phrase (1-a) est une phrase attestée du corpus de l'Est-Républicain (ER), dans laquelle on observe l'ordre SP-SN. L'ordre inverse, présenté dans la phrase (1-b), est également acceptable.

(1) a. *Une manière de montrer* [*au public, essentiellement composé de parents,*] [*les progrès accomplis par les enfants*]. (ER)
b. *Une manière de montrer* [*les progrès accomplis par les enfants*] [*au public, essentiellement composé de parents*].

L'ordre des compléments du verbe ne détermine pas la grammaticalité de la phrase. Il s'agit donc d'un phénomène affecté par des contraintes préférentielles. L'objectif de ce chapitre est de faire le bilan des contraintes préférentielles dont l'impact sur l'ordre des dépendants du verbe a été identifié. Étant donné que, pour le français, nous disposons de peu d'éléments à ce sujet, nous exposerons les contraintes générales postulées pour différentes langues. Cela nous permettra de montrer la diversité des dimensions mises en jeu. Nous nous attacherons également à comprendre en quoi les contraintes présentées agissent sur l'ordre des dépendants du verbe.

Le chapitre sera organisé en six sections. Dans la première, nous présenterons le type de phénomènes étudiés dans les travaux portant sur l'ordre des dépendants du verbe de différentes langues et nous expliciterons, en comparaison avec ces phénomènes, les particularités de la problématique qui nous occupe pour le français. La section suivante mentionnera les contraintes générales postulées dans la littérature comme valables pour toutes les langues. Dans la troisième section, nous traiterons de l'influence de la pronominalité sur l'ordre des dépendants verbaux en anglais, en allemand et en français. La quatrième section sera consacrée aux contraintes liées à la longueur et à la complexité syntaxiques. Dans la section suivante, nous reviendrons sur les aspects lexico-sémantiques qui peuvent avoir une influence sur l'ordre des dépendants du verbe. Enfin, dans la dernière section, nous détaillerons les contraintes relatives à l'organisation du discours.

5.1. Les phénomènes étudiés à travers les langues

À la différence du problème de la position de l'adjectif épithète, les contraintes guidant le choix dans l'ordonnancement des compléments postverbaux en français ont été peu étudiées[1]. À notre connaissance, seuls quatre travaux portent sur la question spécifique de l'ordre de deux compléments sous-catégorisés en français :

1. Frédéric Sabio, spécialiste de l'ordre des mots dans la Macro-Syntaxe, a notamment étudié les structures permettant la réalisation des compléments dans le domaine préverbal (Sabio, 2006, 2007). Cependant, à notre connaissance, il n'a pas étudié les problèmes touchant à l'ordre des compléments dans le domaine postverbal.

Blinkenberg (1928), Berrendonner (1987), Schmitt (1987a,b) et Abeillé & Godard (2000, 2004, 2006). Ces travaux portent sur des observations de phrases attestées et sur des jugements de grammaticalité et d'acceptabilité. Aucun ne fait mention d'études exhaustives sur des données de corpus, ni d'études expérimentales.

Par contraste, il existe un grand nombre de travaux traitant de l'ordre des constituants dans le domaine verbal pour d'autres langues que le français, notamment pour l'anglais. Dans les sections 1.3 et 2.2.2.4 des chapitres 1 et 2, nous avons déjà abordé le problème de l'alternance dative dans cette langue. D'autres phénomènes de l'anglais impliquent la linéarisation des dépendants du verbe : le *Heavy NP Shift* (2), la construction verbe-particule (3) et l'ordre relatif des SP multiples (4) (exemples tirés de Wasow & Arnold, 2003).

(2) *Heavy NP Shift* (HNPS)
   a. *We take [too many dubious idealizations] for granted*
   b. *We take for granted [too many dubious idealizations]*

(3) la construction verbe-particule
   a. *We figure the problem out*
   b. *We figure out the problem*

(4) SP multiples
   a. *Pat talked to Chris about Sandy*
   b. *Pat talked about Sandy to Chris*

Dans les autres langues étudiées, les phénomènes impliquent en général le sujet et un objet du verbe. En allemand, l'ordre des mots connaît une plus grande liberté qu'en anglais ou en français. Les linguistes se sont intéressés à l'ordre relatif du sujet et de l'objet direct (accusatif) ou indirect (datif), dans les propositions principales et dans les propositions subordonnées (exemples tirés de Bader & Häussler, 2010).

(5) Ordre sujet / objet direct - proposition principale
   a. *[Der Direktor]$_{SUJ}$ wird [dieses Buch]$_{OD}$ kaufen*
      le.NOM directeur AUX.FUT ce.ACC livre acheter
      Le directeur va acheter ce livre
   b. *[Dieses Buch]$_{OD}$ wird [der Direktor]$_{SUJ}$ kaufen*
      ce.ACC livre AUX.FUT le.NOM directeur acheter
      Le directeur va acheter ce livre

(6) Ordre sujet / objet direct - proposition subordonnée
   a. *[…] dass [der Direktor]$_{SUJ}$ [dieses Buch]$_{OBJ}$ gelesen hat*
      que le.NOM directeur ce.ACC livre lu a
      …que le directeur a lu ce livre
   b. *[…] dass [dieses Buch]$_{OBJ}$ [der Direktor]$_{SUJ}$ gelesen hat*
      que ce.ACC livre le.NOM directeur lu a
      …que le directeur a lu ce livre

En ce qui concerne l'espagnol, Prat-Sala & Branigan (2000) ont étudié, par exemple, l'influence du caractère donné ou nouveau du référent sur le choix entre voix active et voix passive. Ces auteurs ont également examiné l'effet de ce facteur sur la production de l'ordre SVO, ou OVS avec dislocation à gauche.

(7) a. *El tren atropell-ó a la mujer*
le train renverser-PRET.3SG A la femme
Le train a renversé la femme (SVO - voix active)
b. *La mujer fué atropellada por el tren*
la femme être.PRET.3SG renversé.FEM par le train
La femme a été renversée par le train (SVO - voix passive)
c. *A la mujer la atropell-ó el tren*
A la femme pron.ACC renverser-PRET.3SG la train
Le train a renversé la femme (OVS avec dislocation - voix active)

Enfin, Tanaka *et al.* (2011) se sont intéressés à l'ordre du sujet et de l'objet direct dans des phrases actives, ainsi qu'à l'ordre sujet et objet oblique dans les phrases passives, en japonais.

(8) Voix active
a. *[booto-ga]$_{SUJ}$ [ryoshi-o]$_{OBJ}$ hakonda*
bateau-NOM pêcheur-ACC porter-PSÉ
le bateau portait le pêcheur
b. *[ryoshi-o]$_{OBJ}$ [booto-ga]$_{SUJ}$ hakonda*
pêcheur-ACC bateau-NOM porter-PSÉ
le bateau portait le pêcheur

(9) Voix passive
a. *[ryoshi-o]$_{SUJ}$ [booto-niyotte]$_{OBL}$ hakobareta*
pêcheur-NOM bateau-OBL porter-PASSIF
Le pêcheur était porté par le bateau
b. *[booto-niyotte]$_{OBL}$ [ryoshi-o]$_{SUJ}$ hakobareta*
bateau-OBL pêcheur-NOM porter-PASSIF

L'objet des travaux concernant l'ordre des dépendants du verbe, englobe l'étude du sujet, dans de nombreuses langues. De plus, une partie de ces travaux s'intéresse à l'interaction entre ordre des constituants et assignation des fonctions grammaticales (alternance dative, alternance actif/passif).

Dans ce travail, nous nous concentrons sur l'ordre des compléments postverbaux en français. Il est important de noter que ce phénomène se distingue d'une partie de ceux cités précédemment. Premièrement, nous étudions l'ordre de deux dépendants du verbe, en excluant son sujet. Deuxièmement, le phénomène étudié ne met pas en jeu une alternance de construction. À la différence de l'alternance dative, le phénomène qui nous intéresse ne concerne pas une différence d'appariement entre rôles sémantiques et fonctions grammaticales. Nous nous situons donc dans une pro-

blématique stricte de linéarisation. Enfin, nous avons limité notre objet d'étude aux constituants exclusivement postverbaux.

5.2. Contraintes générales

Les contraintes proposées dans la littérature ont vocation à être valables pour toutes les langues. L'objectif de ces travaux est de décrire l'alignement harmonique des arguments du verbe, c'est-à-dire, de déterminer quelles propriétés grammaticales, sémantiques et discursives convergent vers quelle position relative. Dans les langues étudiées, de grandes tendances s'observent : les référents animés précèdent les référents inanimés, les éléments pronominaux précèdent les non-pronominaux, les constituants définis précèdent les constituants indéfinis. En ce qui concerne le poids, les langues se distinguent selon qu'elles présentent un ordre Objet-Verbe ou Verbe-Objet. Hawkins (1994) a montré que, dans les langues à ordre Verbe - Objet telles que le français ou l'anglais, les constituants courts ont tendance à précéder les constituants plus longs. Inversement, dans les langues à ordre Objet - Verbe, telles que le japonais, les constituants les plus longs ont tendance à précéder les plus courts. Ces tendances sont récapitulées dans la table 5.1 ($\prec$ signifie *tend à précéder*).

Toutes les langues :	référent animé	$\prec$	référent inanimé
	pronominal	$\prec$	non-pronominal
	défini	$\prec$	indéfini
Langues VO :	court	$\prec$	long
Langues OV :	long	$\prec$	court

Table 5.1.: Tendances générales à travers les langues dans l'ordre des constituants ($\prec$ signifie *tend à précéder*)

Pour aborder l'ensemble de contraintes mises à jour par l'étude de différentes langues, nous organiserons l'exposé en suivant les hiérarchies proposées par Bader & Häussler (2010, p. 720), à la suite de Siewierska (1993, p. 831). Ces hiérarchies s'appuient sur les travaux d'Allan (1987) qui répertorie sept types de hiérarchies pertinentes pour expliquer l'ordre des conjoints dans une coordination. Les hiérarchies affectant la linéarisation des constituants sont les suivantes :

(10) Hiérarchies de poids (langues VO)
   a. structurellement simple $\prec$ structurellement complexe
   b. court $\prec$ long

(11) Hiérarchies lexico-sémantiques
   a. Hiérarchie de personne :
   1re personne ≺ 2e personne ≺ 3e personne humain ≺ animaux ≺ matière non-organique ≺ abstraits
   b. Hiérarchie de rôles sémantiques :
   agent ≺ patient ≺ destinataire ≺ bénéfactif ≺ instrumental ≺ spatial ≺ temporel

(12) Hiérarchies relatives au discours
   a. plus familier ≺ moins familier
   b. référent donné ≺ référent nouveau
   c. défini ≺ indéfini
   d. référentiel ≺ non-référentiel

Avant de détailler les différentes hiérarchies, nous présentons l'influence du caractère pronominal sur l'ordre de mots.

5.3. Pronominalité

Le caractère pronominal d'un constituant constitue un facteur influençant très fortement l'ordre des mots. En anglais et en allemand, la présence d'un pronom détermine très largement l'ordre choisi. Par exemple, en allemand, dans une proposition principale contenant un objet accusatif et un objet datif après le verbe, si l'un des deux objets est pronominal, il a tendance à apparaître avant l'autre objet. Dans l'exemple (13), le pronom datif *ihm* est préférentiellement placé directement après le verbe, tandis que dans l'exemple (14), c'est le pronom accusatif *es* qui a tendance à être adjacent au verbe.

(13) Pronom datif
   a. *Der alte Mann hat* ***ihm*** *das Buch geschenkt*
   le vieux homme a 3SG.DAT le livre offert
   Le vieil homme lui a offert le livre

(14) Pronom accusatif
   a. *Der alte Mann hat* ***es*** *seinem Sohn geschenkt*
   le vieux homme a 3SG.ACC son-DAT fils offert
   Le vieil homme l'a offert à son fils

Dans les travaux sur l'anglais et l'allemand, seuls les pronoms personnels définis sont considérés comme pronominaux. Cela signifie que les constituants pronominaux regroupent plusieurs propriétés favorisant leur apparition comme premier constituant, et ce aux trois niveaux présentés dans le classement donné ci-dessus. Premièrement les pronoms sont courts, généralement monosyllabiques, parfois bisyllabiques, et les

constituants à tête pronominale ne sont pas complexes. Deuxièmement, les pronoms sont les réalisations privilégiées des éléments apparaissant en tête de la hiérarchie de personne : 1re, 2e et 3e personne. Troisièmement, ils sont définis et donnés soit par le contexte linguistique précédent (emploi anaphorique), soit par le contexte non-linguistique (emploi déictique). Le pronom porte donc un faisceau de propriétés favorisant l'ordre pronominal - non-pronominal.

On pourrait se demander si l'un des facteurs prime sur les autres et s'il explique à lui seul la préférence pour un ordre. Cependant, les travaux de Bresnan *et al.* (2007) et Gries (2003a), par exemple, ont montré que dans l'alternance dative le caractère pronominal a une influence non réductible à son poids ou à son statut informationnel.

5.3.1. Pour le français

Étant donné que, dans la majeure partie des cas, la pronominalisation des compléments du verbe implique l'utilisation de clitiques préposés au verbe, le nombre d'occurrences de pronoms dans le domaine postverbal est réduit. De plus, dans les études citées précédemment, seuls les pronoms personnels définis sont pris en compte. Cela réduit donc le possibilité d'action de ce facteur aux SP du type *en lui* ou *à moi*. On peut donc émettre l'hypothèse que l'influence du facteur relatif au caractère pronominal est beaucoup moins important en français qu'en allemand ou en anglais.

5.4. Hiérarchies de poids

Hiérarchies de poids (langues VO) :

a) structurellement simple $\prec$ structurellement complexe

b) court $\prec$ long

Le poids est une notion fondamentale dans les phénomènes d'ordre des mots. Rappelons qu'il peut être envisagé de deux façons : soit du point de vue de la longueur, en nombre de mots ou de syllabes, soit du point de vue de la complexité, en nombre de syntagmes, par exemple.

La notion de poids a d'abord été formulée en termes de longueur par Behaghel (1909), qui, à partir de l'observation de l'allemand, du latin et du grec, a remarqué la tendance à placer les éléments les plus courts avant les plus longs.

> « *Ainsi se crée, de manière inconsciente dans les langues, un sentiment rythmique bizarre, la tendance à aller de l'élément le plus court à l'élément le plus long…tendance que je voudrais…nommer* ***la loi des éléments croissants*** »[2]

Hawkins (1994) a documenté la généralisation de Behaghel avec des études sur différentes langues, menées à partir de corpus. Il a notamment mis à jour la tendance

2. « *So bildet sich unbewußt in den Sprachen ein eigenartiges rhythmisches Gefühl, die Neigung, vom kürzeren zum längeren Glied überzugehen…was ich…als das Gesetz der wachsenden Glieder bezeichnen möchte* », (Behaghel, 1909, 139)

« miroir », c'est-à-dire *long précède court*, dans certaines langues à tête finale, tel que le japonais.

Le poids a aussi été envisagé en termes de complexité, notamment chez Chomsky (1975). Cet auteur remarque qu'un objet direct comportant une proposition relative aura tendance à apparaître en dernier, même s'il est court :

> « *Il est intéressant de remarquer que ce n'est apparemment pas la longueur en nombre de mots de l'objet qui détermine le naturel de la transformation, mais plutôt, d'une certaine façon, sa complexité. Ainsi, "they brought all the leader of the riot in" semble plus naturel que "they brought the man I saw in". Ce dernier, bien que plus court, est plus complexe.* » [3]

Wasow (1997, 2002) a étudié plus précisément la notion de poids pour l'anglais, à partir de données de corpus. Il en conclut que cette notion doit être définie de façon relative et graduée, que ce soit en termes de longueur ou en termes de complexité. Il montre que la relation entre poids et pourcentage de HNPS [4] est graduelle : par exemple, pour un SN contenant 3 noeuds syntagmatiques, il observe moins de 10% de HNPS, tandis que pour 6 noeuds, il en trouve environ 40%, et près de 80% quand le nombre de noeuds est supérieur ou égal à sept. Wasow rejette donc les définitions absolues du poids, telles que celle de Erdmann (1988) : « *Considérer un groupe nominal comme lourd signifie soit que deux ou plusieurs groupes nominaux sont coordonnés, soit que le nom tête du groupe nominal est post-modifié par un syntagme ou une proposition.* » [5]

L'auteur montre également que les définitions de poids, en termes de longueur ou de complexité, rendent mieux compte des données des corpus si elles sont considérées d'un point de vue relatif. En effet, la prédiction de l'ordre attesté en corpus est meilleure lorsque l'on compare la longueur des deux constituants qui entrent en jeu, que lorsque l'on estime isolément le poids de chaque constituant. Par exemple, pour un constituant de 5 mots, la tendance varie selon que l'autre constituant contient 2 mots ou plus de 10 mots. Ainsi, dans l'exemple (15), le SP apparaît naturellement en dernier lorsque le SN ne compte que deux mots, tandis qu'il semble plus naturel de le placer à côté du verbe dans le cas où le SN fait plus de 10 mots.

(15) a. *Paul offrira un livre [à son plus jeune fils].*
b. *Paul offrira [à son plus jeune fils] un très bel ouvrage illustré sur la faune et la flore des Pyrénées.*

La définition de la notion de poids doit donc être **graduée** et **relative** : *soient les constituants A et B, plus le constituant A est lourd par rapport au constituant B, plus*

3. « *It is interesting to note that it is apparently not the length in words of the object that determines the naturalness of the transformation, but, rather, in some sense, its complexity. Thus "they brought all the leader of the riot in" seems more natural than "they brought the man I saw in". The latter, though shorter, is more complex* » (Chomsky, 1975, p. 477)

4. *Heavy NP Shift.*

5. « *Counting a nominal group as heavy means either that two or more nominal groups […] are coordinated […], or that the head noun of a nominal group is postmodified by a phrase or clause.* », (Erdmann, 1988, p. 328)

on aura tendance à choisir l'ordre B A.

Toujours à partir de données de corpus, Wasow (1997) a comparé les mesures de poids en termes de complexité et de longueur. En ce qui concerne la longueur, il considère le nombre de mots, à l'image de ce que propose Hawkins (1990) [6]. Quant à la complexité, il examine deux mesures : le nombre de noeuds syntaxiques dominés par le constituant, inspiré de Hawkins (1994), et le nombre de noeuds syntagmatiques contenu par le constituant, d'après Rickford *et al.* (1995). Il observe que ces mesures ont une capacité de prédiction équivalente sur les données de corpus : l'une ne semble pas meilleure que les deux autres. Cela s'explique notamment par le fait que ces trois mesures sont extrêmement corrélées. Sachant qu'un coefficient de corrélation égal à 1 signifie une corrélation parfaite des données [7], le tableau 5.2 montre que le nombre de mots, le nombre de noeuds et le nombre de noeuds syntagmatiques sont des mesures quasi équivalentes pour le HNPS, l'alternance dative (AD) et la construction Verbe - Particule (VP).

	HNPS	AD	VP
Mots et noeuds	.94	.96	.99
Mots et noeuds syntagmatiques	.96	.97	.95
Noeuds et noeuds syntagmatiques	.94	.96	.98

TABLE 5.2.: Coefficients de corrélation pour les mesures de poids (AD = Alternance dative ; VP = construction Verbe-Particule) (Wasow, 1997, p. 93).

Les syntagmes qui contiennent beaucoup de mots ont tendance à avoir des structures plus complexes, et donc plus de noeuds et plus de noeuds syntagmatiques. Cette étude sur corpus ne permet pas de décider quelle mesure constitue le meilleur prédicteur pour les phénomènes d'ordre en anglais. Ces observations justifient l'utilisation de la longueur des constituants en nombre de mots comme un paramètre dans les études sur les facteurs influençant l'ordre en anglais.

Dans leur recherche sur la notion de poids, Wasow & Arnold (2003) ont montré que l'ordonnancement des constituants est sensible à la complexité, indépendamment de la longueur. Ils ont d'abord utilisé un questionnaire portant sur le HNPS, l'alternance dative et la construction Verbe - Particule. Le questionnaire contenait des phrases dont les syntagmes susceptibles d'alterner avaient la même longueur, mais différaient quant à leur complexité. Les auteurs ont demandé à 88 sujets de juger l'acceptabilité sur une échelle de 1 (= complètement inacceptable) à 4 (= complètement acceptable). Les résultats montrent que les phrases contenant un syntagme complexe en position finale tendent à recevoir une note significativement meilleure. De même, pour les phrases comportant un constituant simple, la note est significativement meilleure lorsque le constituant en question apparaît dans une position non-finale. Cela signi-

6. Wasow (1997) ne propose pas d'étudier la longueur en nombre de syllabes.

7. Wasow (1997) ne précise pas quel coefficient de corrélation il a utilisé. On peut supposer qu'il s'agit d'un ρ de Spearman ou d'un ρ de Pearson.

fie qu'indépendamment de la longueur, la complexité des constituants influence les jugements des locuteurs.

Wasow & Arnold (2003) ont également étudié la longueur et la complexité à partir de données du corpus portant sur le HNPS, l'alternance dative et la construction Verbe - Particule. La longueur a été évaluée en nombre de mots. La mesure de la complexité a été produite à partir de deux catégories : *complex* (modification après la tête du syntagme) ou *simple* (pas de modification après la tête). Les données indiquent que les facteurs complexité et longueur sont de meilleurs prédicteurs lorsqu'ils sont combinés pour le HNPS et pour l'alternance dative. En revanche, la complexité ne semble pas être pertinente dans le cas de la construction Verbe - Particule. Les auteurs suggèrent que, la particule étant particulièrement légère dans cette construction, les SN de plus de 3 mots ont tendance à apparaître en position finale, quelle que soit leur complexité.

Une fois la notion de poids définie et identifiée pour chaque phénomène, se pose la question de l'effet du poids grammatical : pourquoi la longueur a-t-elle une influence sur l'ordre des constituants ? De façon générale, il existe deux points de vue pour répondre à cette question : celui du traitement des phrases par l'interlocuteur, que nous détaillerons à partir des travaux de Hawkins (1994) et celui de la planification par le locuteur, défendu notamment par Wasow (2002).

Hawkins (1994) rend compte des phénomènes de poids dans l'ordonnancement des constituants en introduisant des contraintes liées au traitement (*processing*) des phrases. Il pose le principe *Early Immediate Constituent* (EIC) selon lequel : « *L'analyseur syntaxique humain préfère les ordres linéaires qui maximisent les ratios IC-to-non-IC des domaines d'identification des constituants* » [8]. Pour comprendre ce principe, il faut définir le ratio *IC-to-non-IC* et le domaine d'identification du constituant.

Le *domaine d'identification des constituants* (DIC) correspond à la séquence de mots minimale permettant d'identifier tous les constituants immédiats du SV. Ainsi, dans un SV composé de 3 constituants, le DIC s'étend du premier mot du SV jusqu'au premier mot du troisième constituant. Le DIC du SV en (16) est *gave to Mary the*. En effet, le premier mot du SN (*the*) permet de déterminer immédiatement le noeud syntagmatique auquel correspond le dernier constituant. Selon les termes de Hawkins, le noeud syntagmatique est immédiatement *construit* au-dessus de l'article *the* : ce mot est donc suffisant pour identifier le dernier constituant.

(16) I [SV gave [SP to Mary] [SN the valuable book that was extremely difficult to find]].

Le ratio *IC-to-non-IC* correspond au rapport entre le nombre de constituants immédiats dans le DIC et le nombre de constituants non-immédiats. Pour calculer ce ratio chaque mot du SV est associé à une fraction, dans laquelle le dénominateur correspond à la place du mot dans le SV et le numérateur renvoie au nombre de

8. « *The human parser prefers linear orders that maximize the IC-to-non-IC ratios of constituent recognition domains* »

constituants immédiats identifiables à ce moment-là dans le SV. Le ratio *IC-to-non-IC* correspond à la moyenne des fractions dans le DIC. Pour les phrases (17) et (18), nous exemplifions le calcul du ratio. Théoriquement, plus le ratio est élevé pour une phrase donnée, plus l'ordre attesté dans cette phrase est préféré.

(17) *I [$_{SV}$ gave [$_{SN}$ the valuable book that was extremely difficult to find] [$_{SP}$*
1/1 2/2 2/3 2/4 2/5 2/6 2/7 2/8 2/9 2/10
to Mary]].
3/11 = 46,6%

(18) *I [$_{SV}$ gave [$_{SP}$ to Mary] [$_{SN}$ the valuable book that was extremely difficult*
1/1 2/2 2/3 3/4
to find]].
= 85,4%

D'après le principe EIC, la phrase (18) est préférée par l'analyseur syntaxique humain. En effet, il semble que cette phrase soit plus naturelle que la phrase (17). Ce principe se veut général, c'est-à-dire applicable à toutes les langues. Cette notion très précise sous-tend l'idée selon laquelle la complexité syntaxique des constituants, ou plus précisément le nombre de noeuds, est le principal facteur expliquant les phénomènes d'ordre liés au poids.

Le postulat servant de base au principe EIC, ainsi qu'à toute la théorie de Hawkins (1994), présente la facilitation du traitement des phrases comme la principale utilité des phénomènes d'ordre liés au poids.

Wasow (2002, p. 42-45) remet cette analyse en cause et lui oppose l'hypothèse selon laquelle l'ordonnancement des constituants renvoie essentiellement à la planification et à la production, autrement dit, est à comprendre du point de vue du locuteur. Il estime que le principal problème posé par l'hypothèse de Hawkins (1994) est que, si le locuteur calcule un ratio du type de *IC-to-non-IC*, il doit planifier la phrase entière avant de choisir l'ordre. Or, il semble que les locuteurs aient tendance à construire les phrases au fur et à mesure de leur production, comme en témoignent par exemple les indices du type disfluences, répétitions, faux départs (Clark, 1994, 1996; Clark & Wasow, 1998). L'un des arguments de Wasow (2002) repose sur l'analyse de données de corpus. L'étude de l'auteur porte sur le HNPS, dans le cas où le SP forme avec le verbe une séquence conventionalisée, ou collocation selon les termes de l'auteur. Cette séquence peut être sémantiquement transparente comme dans *bring to an end*, ou opaque, comme dans *take into account*. L'hypothèse de l'auteur est que la présence d'un HNPS (ordre V SP SN) avec une collocation, qu'elle soit opaque ou non, est bénéfique au locuteur, dans la mesure où produire la séquence collocative en premier lui permet de réduire sa charge mémorielle et d'avoir plus de temps pour planifier et produire le SN objet. Par contraste, la présence d'un HNPS avec une collocation n'est bénéfique à l'interlocuteur que dans le cas d'une collocation opaque. En effet, en ayant connaissance de la dimension collocative du couple verbe/SP, l'interlocuteur peut assigner la bonne interprétation au verbe. Dans le cas des collocations sémantiquement transparentes, l'interprétation de l'interlocuteur n'est pas facilitée par la présence du

HNPS. Ainsi, si l'ordre des constituants est établi dans la perspective du locuteur, on observera une proportion d'HNPS plus importante quand le verbe et le SP forment une collocation, que lorsqu'il n'y a pas de collocation ; si l'ordre est établi dans la perspective de l'interlocuteur, on observera une proportion d'HNPS plus importante pour les collocations opaques que pour les collocations transparentes. Les données recueillies dans le corpus *Aligned-Hansard* sont compatibles avec les deux hypothèses. De façon générale, il y a plus de HNPS (54%) en présence d'une collocation qu'en son absence (15%). L'auteur constate donc un effet de la collocation sur la production de HNPS. Par ailleurs, les collocations opaques présentent plus fréquemment le HNPS (60%) que les collocations transparentes (47%). Pour les deux types de collocation, Wasow observe un effet supérieur de la collocation opaque. Il estime néanmoins que la différence d'effet entre collocation opaque et collocation transparente sur l'ordre des syntagmes est moins significative que celle produite par la distinction générale collocation/non-collocation. Cela montre, selon l'auteur, que l'effet de la planification de l'énoncé est plus important que celui de l'analyse de l'interlocuteur.

Wasow conclut que le locuteur a intérêt à placer les éléments les plus lourds à la fin, dans la mesure où « *postposer les constituants difficiles et garder des options ouvertes aussi longtemps que possible facilite la planification pendant la production d'un énoncé* » [9].

5.4.1. Pour le français

Dans ses quelques pages consacrées à l'ordre relatif du complément d'objet direct et indirect, Blinkenberg (1928) évoque l'idée selon laquelle les compléments sont ordonnés par longueur croissante. Abeillé & Godard (2004, 2006) considèrent, plus généralement, que les constituants postverbaux s'organisent selon leur poids. Pour rendre compte de cela, les auteurs proposent une notion de poids à trois valeurs : *léger*, *moyen* et *lourd*. Les mots du lexique peuvent être *légers* ou *moyens*, et la valeur *lourde* est réservée aux syntagmes présentant une certaine complexité syntaxique. En s'appuyant sur l'exemple (19), les auteurs estiment que la lourdeur des syntagmes ne concerne pas la simple longueur des constituants en syllabes.

(19) a. *Jean présentera [M. Konstantin Rastapopoulos] à Marie*
b. *Jean présentera à Marie [M. Konstantin Rastapopoulos]*

De plus, Abeillé & Godard (2004) proposent une notion de légèreté à vocation universelle et définie comme un type de déficience syntaxique. Les mots et les syntagmes légers présentent des propriétés syntaxiques spécifiques : ils sont dépourvus de mobilité et de possibilité d'extraction. Cependant, à la différence des formes faibles, ils peuvent être modifiés ou coordonnés. Grâce à cette notion appliquée au français, les auteurs rendent compte du fait que, dans le domaine verbal, les noms sans déterminant précèdent les autres compléments, comme cela est illustré en (20) et

9. « *postponing difficult constituents and keeping options open as long as possible facilitates planning during utterance production* » (Wasow, 2002, p. 56).

(21) (exemples de Abeillé & Godard (2004, p. 5), accompagnés des jugements des auteurs) [10].

(20) a. *Cet endroit fait [peur] aux enfants*
b. **Cet endroit fait aux enfants [peur]*

(21) a. *Le Président rendra [hommage] aux victimes*
b. **Le Président rendra aux victimes [hommage]*

La valeur *légère* est attribuée aux nom nus et une contrainte d'ordre stipule que les constituants légers doivent précéder les constituants non-légers.

L'un des objectifs de notre travail sur corpus sera de déterminer plus précisément à quoi correspond la notion de lourdeur et si elle est comparable à celle rencontrée en anglais (Wasow, 1997). Nous pourrons également observer si les données sont conformes aux généralisations de Abeillé & Godard (2004) en ce qui concerne la légèreté et les noms nus.

5.5. Hiérarchies lexico-sémantiques

Dans cette section, nous détaillerons les hiérarchies de personne et de rôle sémantique et nous les exemplifierons à partir de travaux sur l'anglais et sur l'allemand. À la suite de la hiérarchie des rôles sémantiques, nous discuterons du lien entre le verbe et l'un des deux constituants sous-catégorisés. En effet, les rôles sémantiques touchent à une facette du rôle du verbe qu'il semble intéressant de compléter par une autre : la connexion possible entre le verbe et d'autres éléments.

5.5.1. Hiérarchie de personne

(22) 1re personne $\prec$ 2e personne $\prec$ 3e personne humain $\prec$ animaux $\prec$ matière non-organique $\prec$ abstraits

La hiérarchie de personne regroupe deux dimensions : la personne grammaticale et le caractère animé. Il s'agit en fait de subdiviser les référents humains en trois sous-groupes formés par les personnes grammaticales. Cette subdivision est pertinente dans le cas où les pronoms personnels sont impliqués. Dans la section 5.3, nous avons commenté le cas des pronoms qui ne relève pas seulement de la hiérarchie de personne, mais également des hiérarchies de poids et de celles relatives au discours. Ici, nous nous concentrons sur le caractère animé, à savoir la deuxième partie de la hiérarchie : '*humain* $\prec$ *animaux* $\prec$ *matières non-organiques* $\prec$ *abstraits*'. Il existe plusieurs hiérarchies possibles se rapportant au caractère animé (Garretson, 2004; Yamamoto, 1999). Cependant, il est admis que, de façon générale, la hiérarchie s'organise de la façon suivante :

10. La contrainte de légèreté permet plus largement de rendre compte des phénomènes d'ordre en français tels que l'ordre relatif du nom et de l'adjectif dans le SN, la position des adverbes dans le domaine verbal, ainsi que des pronoms interrogatifs et des pronoms quantifieurs.

(23) humains ≺ autres animés ≺ inanimés

Cette hiérarchie, parfois simplifiée en *animé* vs. *inanimé*, s'exprime directement au niveau de la grammaire dans certaines langues. Par exemple, en espagnol, l'objet est accompagné du marqueur *a*, lorsqu'il renvoie à une entité animé et référentielle (Pensado, 1995; Torrego, 1999, parmi d'autres)[11]. Dans l'exemple (24-a) où l'objet direct réfère à un être animé, le marqueur *a* est obligatoire, alors que sa présence avec un objet inanimé rend la phrase (24-b) agrammaticale.

(24) a. *Pedro ha visto* ***a*** *su madre (*Pedro ha visto su madre)*
Pedro a vu A sa mère
Pierre a vu sa mère
b. *Pedro ha visto su libro (*Pedro ha visto* ***a*** *su libro)*
Pedro a vu le livre
Pedro a vu le livre

La hiérarchie du caractère animé s'exprime également au niveau de l'ordre des mots. Nous avons vu par exemple, dans le chapitre 1, qu'en sesotho (Morolong & Hyman, 1977), l'ordonnancement des compléments des verbes ditransitifs est soumis à une contrainte catégorique reposant sur l'opposition *humain/non-humain* : s'il y a une asymétrie entre la nature des référents des deux arguments non-sujet (thème et bénéficiaire), le constituant renvoyant à un référent humain doit apparaître obligatoirement à côté du verbe.

Le caractère animé est également impliqué dans l'assignation des fonctions grammaticales dans la phrase. Par exemple, Bock *et al.* (1992) montrent, à partir d'expériences psycholinguistiques, que les locuteurs de l'anglais ont tendance à produire des phrases à la voix passive lorsqu'ils décrivent une action où le patient est animé et l'agent inanimé. De cette façon, les locuteurs tendent à faire correspondre la fonction grammaticale sujet avec le référent animé.

Afin de préciser la manière dont le caractère animé du référent peut exercer une influence sur l'ordre des mots, nous reprenons l'article de Branigan *et al.* (2008) qui expose clairement les enjeux du caractère animé du point de vue de la production. Pour produire des énoncés bien formés, il faut traiter beaucoup d'informations de niveaux différents (phonétique, prosodie, syntaxe, sémantique…) et de façon simultanée. Pour rendre compte de cette activité très intense lors de la production, les psycholinguistes postulent généralement que le traitement des différents aspects de la production d'un énoncé est incrémental. Chaque niveau est mis à jour et construit de façon parallèle, au fur et à mesure que le locuteur parle. Une fois l'hypothèse d'incrémentalité de la production posée, l'accessibilité de l'information apparaît comme un élément fondamental dans la production des énoncés : un élément facilement accessible est traité avant un élément qui l'est moins. L'accessibilité d'une information est en partie influencée par l'accessibilité conceptuelle telle qu'elle est définie par Bock & Warren

11. Le marquage différentiel de l'objet implique d'autres facteurs relatifs aux propriétés syntaxiques et sémantiques du verbe, voir par exemple von Heusinger & Kaiser (2011).

(1985) : « *la facilité avec laquelle la représentation mentale d'un référent potentiel peut être activée ou récupérée dans la mémoire* » [12]. Les entités animées ont un haut degré d'accessibilité conceptuelle. Elles sont donc généralement plus faciles à récupérer ou à activer que des entités inanimées. Cela signifie que les entités animées ont tendance à être traitées avant les autres et, donc à être produites en premier. Une telle hypothèse implique que le caractère animé des référents est un facteur universel, influençant l'ordre des mots dans toutes les langues, quand aucune autre contrainte catégorique n'intervient.

Un débat psycholinguistique sur le rôle du caractère animé dans la production oppose les tenants de l'hypothèse *indirecte* à ceux de l'hypothèse *directe* [13]. Selon l'hypothèse *indirecte*, le caractère animé influence de manière directe l'assignation des fonctions grammaticales et de façon indirecte l'ordre des mots, dans la mesure où les fonctions grammaticales associées aux référents animés apparaissent généralement en premier dans l'ordre linéaire. L'hypothèse *directe* stipule que le caractère animé influence l'assignation des fonctions grammaticales et l'ordre des constituants de façon indépendante.

L'hypothèse *indirecte* a notamment été défendue par McDonald *et al.* (1993). Leur travail s'appuie sur des expériences de production et de compréhension, afin d'étudier le rôle du caractère animé des noms dans l'ordre des mots et dans l'assignation des fonctions grammaticales. Les auteurs montrent que les noms animés tendent à être utilisés comme sujet, ce qui va dans le sens de l'idée selon laquelle les référents animés ont tendance à se voir assigner les rôles grammaticaux les plus hauts dans la hiérachie des relations grammaticales '*sujet* $\prec$ *objet direct*$\prec$ *objet indirect et oblique*' (Bock, 1987). En anglais, la hiérarchie des relations grammaticales correspond généralement à l'ordre linéaire des constituants, ce qui permet de dire que l'assignation des fonctions grammaticales renforce la présence des référents animés avant les référents inanimés. De plus, McDonald *et al.* (1993) observent que, dans une coordination de deux noms, le nom animé a tendance à précéder l'inanimé, mais ce uniquement en séquence isolée. Lorsque la coordination est insérée dans un énoncé complet, il n'y a pas d'effet du caractère animé sur l'ordre des deux noms. Les auteurs en concluent que, quand deux noms partagent la même fonction grammaticale, le rôle du caractère animé est neutralisé.

Comme le font remarquer Branigan *et al.* (2008), l'ordre des constituants en anglais est relativement fixe, une fois les fonctions grammaticales assignées. Ainsi, cette langue ne semble pas être le lieu privilégié pour l'observation du rôle du caractère animé sur l'ordre des constituants. C'est pourquoi l'étude de langues à ordre des mots plus libre est pertinente pour traiter de cette question. Les travaux de Branigan & Feleki (1999) sur le grec, de Tanaka *et al.* (2011) sur le japonais et de Kempen & Harbusch (2004) sur l'allemand, défendent l'hypothèse *directe*, selon laquelle le caractère animé influence non seulement l'assignation des fonctions grammaticales, mais aussi

12. « *the ease with which the mental representation of some potential referent can be activated in or retrieved from memory.* », (Bock & Warren, 1985, p. 50)

13. Nous empruntons les expressions *hypothèse directe* et *hypothèse indirecte* à Kempen & Harbusch (2004).

l'ordre des constituants (Branigan *et al.*, 2008).

Le grec est une langue casuelle qui autorise, dans l'ordre des mots, une plus grande liberté qu'une langue configurationnelle telle que l'anglais. Branigan & Feleki (1999) ont étudié l'ordre SVO et l'ordre OVS à partir d'expériences psycholinguistiques. Les sujets de l'expérience devaient répéter des phrases préalablement entendues, qui présentaient alternativement l'ordre SVO et OSV et pour lesquelles les référents du sujet et de l'objet étaient alternativement animés ou inanimés. Les auteurs ont observé que, quelle que soit la fonction grammaticale, les locuteurs du grec ont tendance à produire la phrase entendue sous une forme dans laquelle le référent animé apparaît avant le verbe et le référent non-animé apparaît après le verbe. Cela constitue un argument fort pour soutenir l'hypothèse selon laquelle le caractère animé affecte l'ordre des constituants indépendamment de l'assignation des fonctions grammaticales.

Dans le même ordre d'idées, Tanaka *et al.* (2011) ont étudié l'influence exercée par le caractère animé sur le choix de l'ordre des constituants, en japonais. Pour cette langue, les deux ordres testés étaient SOV ou OSV. Les deux expériences reportées dans Tanaka *et al.* (2011) impliquaient des variations sur le caractère animé du sujet et de l'objet, ainsi que sur la voix utilisée (active ou passive). À l'instar de l'expérience menée sur le grec, les sujets devaient répéter des phrases préalablement entendues et les observations portaient sur l'ordre des constituants et la voix choisie dans les phrases produites. Les effets du caractère animé sur l'ordre des mots sont similaires à ceux dégagés pour le grec : les locuteurs du japonais tendent à placer en position initiale les référents animés, indépendamment de la fonction grammaticale qui leur est assignée.

Ce phénomène a également été identifié en allemand par Kempen & Harbusch (2004), à partir de données de corpus. Ces auteurs ont examiné l'ordre attesté dans des phrases où l'objet direct (OD) est pronominal et le sujet (SUJ) non pronominal. Comme nous l'avons vu dans la partie 5.3, le caractère pronominal d'un constituant favorise très fortement sont apparition en première position en allemand. Cependant, les auteurs observent une asymétrie : lorsque le sujet est inanimé, l'ordre '$\text{OD}_{pro} \prec \text{SUJ}_{non-pro}$' est attesté à 85%, tandis que lorsque le sujet est animé, l'ordre '$\text{OD}_{pro} \prec \text{SUJ}_{non-pro}$' n'est plus attesté qu'à 51%. Cela suggère que, pour le choix de l'ordre, le caractère animé entre en conflit avec la contrainte préférentielle liée à la pronominalité. De même, Kempen & Harbusch ont observé l'ordre relatif du sujet (SUJ) et de l'objet indirect (OI), sachant que la fonction SUJ a tendance à précéder la fonction OI, toute chose égale par ailleurs. Les données du corpus montrent que lorsque l'OI est inanimé, l'ordre attesté correspond à près de 93% à l'ordre 'SUJ $\prec$OI'. En revanche, lorsque l'OI est animé, l'ordre 'SUJ $\prec$ OI' n'est plus observé qu'à 54%. La contrainte relative au caractère animé semble, une fois de plus, entrer en conflit avec la règle qui tend à ordonner les constituants selon leur fonction. Ces deux observations sur corpus constituent donc deux arguments de plus en faveur de l'hypothèse selon laquelle le caractère animé influence directement l'ordre des mots.

Enfin, il est intéressant de noter que le caractère animé est corrélé à la notion de poids. En effet, les éléments humains et animés ont tendance à être plus courts que les éléments non-animés. Hawkins (1994, p. 424) émet l'hypothèse selon laquelle

l'influence du caractère animé n'est que la conséquence de la corrélation entre poids et caractère animé. Cependant, le travail de Rosenbach (2005) s'attache à montrer que cette affirmation n'est pas vraie et que le caractère animé a une influence indépendamment de celle du poids. L'auteur s'appuie sur des données de corpus, ainsi que sur des données expérimentales relatives à la variation génitive en anglais (*the king's palace* vs. *the palace of the king*). Elle montre que les deux types de données vont dans le même sens : le caractère animé ne peut pas être réduit au facteur poids. Un autre point intéressant est la variation de l'importance du caractère animé selon les variétés d'anglais. Dans leur article traitant de l'alternance dative en américain et en néo-zélandais, Bresnan & Hay (2008) montrent que les locuteurs d'anglais néo-zélandais sont plus sensibles au caractère animé dans l'alternance dative que les locuteurs d'anglais américain. Ce résultat suggère que le caractère animé agit de façon subtile et différenciée à travers les langues et leurs variétés.

5.5.2. Hiérarchie des rôles sémantiques

(25) agent $\prec$ patient $\prec$ destinataire $\prec$ bénéficiaire $\prec$ instrumental $\prec$ spatial $\prec$ temporel

La hiérarchie des rôles sémantiques est à mettre en parallèle avec celle des fonctions syntaxiques, selon laquelle les fonctions grammaticales s'organisent de la façon suivante : 'sujet $\prec$ objet direct $\prec$ objet indirect' (Keenan & Comrie, 1977). Les deux hiérarchies permettent de rendre compte de la correspondance générale entre arguments sémantiques et fonctions syntaxiques. Dans la perspective de la linéarisation des compléments du verbe, déterminer l'influence de la hiérarchie des rôles sémantiques est problématique, dans la mesure où il n'existe pas de consensus à propos de cette hiérarchie, comme le prouve le panorama dressé par Levin & Rappaport Hovav (2005, chap. 2).

Cependant, Bader & Häussler (2010) remarquent deux phénomènes intéressants dans le cas de l'ordre des mots en allemand. Premièrement, dans une phrase active comportant un agent sujet (SUJ), un patient objet direct (OD) et un destinataire objet indirect (OI), l'ordre neutre est 'SUJ $\prec$ OI $\prec$ OD', tandis que dans une phrase passive où le patient est promu SUJ et où l'agent n'est pas exprimé, l'ordre neutre est 'OI $\prec$ SUJ'. Pour rendre compte de façon unifiée de l'ordre des constituants à la voix active et à la voix passive, il semble que l'utilisation de la hiérarchie de rôles sémantiques soit pertinente : à la voix passive comme à la voix active, la phrase neutre allemande s'organise selon la hiérarchie de rôles sémantiques suivante : '*agent* $\prec$ *destinataire* $\prec$ *patient*'. Cette idée est récapitulée dans la table 5.3, inspirée de Bader & Häussler (2010, p. 733).

Néanmoins, comme le notent Bader & Häussler, l'ordre '*destinataire* $\prec$ *patient*' ne correspond pas à la hiérarchie proposée en (25), où patient apparaît plus haut que destinataire. Étant donné que la hiérarchie en (25) rend compte de la correspondance entre arguments sémantiques et fonctions syntaxiques, il faudrait postuler deux hiérarchies différentes pour l'allemand : la première pour la correspondance avec

voix active	SUJ	$\prec$	OI	$\prec$	OD
voix passive	(von-SP)	$\prec$	OI	$\prec$	SUJ
	Agent	$\prec$	**Destinataire**	$\prec$	**Patient**

TABLE 5.3.: Organisation de la phrase active et de la phrase passive en allemand

les fonctions grammaticales et la seconde pour la linéarisation des compléments du verbe.

Deuxièmement, certaines classes de verbes préfèrent l'ordre objet - sujet à la voix active. Bader & Häussler (2010) reprennent la classification proposée par Eisenberg (2004) selon laquelle il existe trois classes verbales qui favorisent l'ordre 'objet $\prec$ sujet'. Ces trois classes se caractérisent par un sujet prototypiquement inanimé et un objet (OD ou OI) remplissant un rôle d'expérient, de cause ou de possesseur.

Prendre en considération les rôles sémantiques des arguments du verbe revient plus largement à étudier les verbes et leur sémantique. De façon générale, on observe que chaque verbe a une préférence plus ou moins marquée pour un ordre : l'item verbal spécifique joue un rôle important. Par exemple, pour l'alternance dative Wasow & Arnold (2003, p. 13-15) montrent que certains verbes apparaissent beaucoup plus fréquemment avec la construction à double objet, tandis que d'autres sont beaucoup plus fréquemment suivis d'une construction dative prépositionnelle. Ils ont relevé une centaine d'occurrences de chaque verbe dans le *New York Times* et ont calculé les proportions en fonction des deux constructions possibles : le verbe *give*, par exemple, apparaît à plus de 80% dans une construction à double objet, quand le verbe *sell* est à moins de 20% dans ce même type de construction.

Bresnan *et al.* (2007, p. 21-25) reviennent sur cet aspect dans leur étude sur l'alternance dative. Ils considèrent que le contenu sémantique du verbe influence le statut des référents du thème et du destinataire. Ainsi, bien que les verbes *bring* et *take* appartiennent à la même large classe sémantique de transfert de possession, le statut du destinataire est différent selon le verbe. En effet, le point de vue du locuteur sur le destinataire n'est pas le même pour le verbe *bring* et pour le verbe *take*. Schématiquement, le destinataire de *bring* est situé près du locuteur, tandis que celui de *take* est éloigné du locuteur. Bresnan *et al.* observent, dans leur données extraites du *Switchboard*, que le verbe *bring* a presque trois fois plus de destinataires étiquetés *donnés* que le verbe *take* ; et inversement, le verbe *take* a environ sept fois plus de destinataires étiquetés *non-donnés*[14]. Ces données confirment l'idée selon laquelle la sémantique du verbe a une influence sur le statut même de ses arguments. De plus, un même verbe peut avoir plusieurs usages renvoyant à différents aspects de ses arguments. Par exemple, le verbe *give* peut avoir un usage abstrait (*it gives it some variety*), pour lequel la proportion de destinataire inanimé est plus grande que la moyenne. Ce verbe peut également être employé dans un sens communicatif (*give me your name*). Si tel est le cas, les destinataires sont, dans une large majorité,

14. Nous aborderons les facteurs relatifs à la structure informationnelle, tels que l'opposition *donné* vs. *nouveau* dans la section 5.6

animés. Le type d'argument d'un même verbe a donc tendance à varier selon l'emploi de ce dernier.

Les deux exemples de Bresnan *et al.* (2007) que nous venons de citer indiquent que la sémantique générale du verbe, ainsi que ses emplois plus spécifiques, ont une influence sur le statut et le type de référents des arguments du verbe. Il faut noter que le biais lexical porte sur des dimensions telles que le caractère animé ou le statut informationnel. Ces dimensions sont généralement envisagées comme des facteurs influençant l'ordre des mots (cf. sections 5.5.1 sur le caractère animé et 5.6 sur le statut informationnel). La corrélation entre la sémantique du verbe et ces facteurs pose la question suivante : le statut informationnel et le caractère animé ont-ils un effet sur l'ordre des mots, indépendamment de la sémantique et de l'emploi du verbe ? Autrement dit, ne pourrait-on pas réduire les effets observés au verbe employé ? En ce qui concerne l'alternance dative en anglais, Bresnan *et al.* (2007) montrent que la prise en compte des verbes et de leur emploi en contexte dans une modélisation statistique du phénomène n'élimine pas les effets de variables relatives au caractère animé et au statut informationnel. Cela signifie que le caractère animé et le statut informationnel ont un effet sur le choix de la construction dative, indépendamment du verbe utilisé. Le verbe influence donc les propriétés sémantiques et discursives de ses arguments, mais les propriétés en question gardent une certaine indépendance par rapport à ce verbe.

L'influence du lemme verbal dans l'ordonnancement des arguments du verbe a également été envisagée d'un point de vue plus syntaxique, en lien avec le traitement de la phrase. Wasow (1997) observe sur des données de corpus le comportement de certaines classes de verbe dans l'alternance dative et dans le HNPS. Premièrement, l'auteur distingue les verbes pouvant sous-catégoriser une proposition subordonnée ou infinitive en plus d'un $\text{SN}_{destinaire}$ (V_s, par exemple *tell Mary that it was raining*) et ceux ne pouvant pas sous-catégoriser de subordonnée (V_n, par exemple *give*). Les données extraites des corpus *Brown* et *Switchboard* présentent une proportion significativement plus élevée de construction à double objet (V $\text{SN}_{destinataire}$ SN_{theme}) pour les V_s que pour les V_n. Il semble donc que la possibilité pour un V_s de sous-catégoriser une proposition finie ou non finie ait une influence sur le choix de la construction dans l'alternance dative. Deuxièmement, Wasow (1997) a observé l'influence des cadres de sous-catégorisation du verbe sur le phénomène du HNPS. L'auteur a créé deux groupes de verbes : ceux qui sous-catégorisent obligatoirement un SN objet (V_t, par exemple *bring*) et ceux qui peuvent sous-catégoriser un SP seul (V_p, par exemple *write*). Les données extraites des corpus *Brown* et *Switchboard* révèlent que les V_p présentent une proportion plus importante de HNPS que les V_t. Autrement dit, on observe plus fréquemment l'ordre V SP SN, avec un verbe ne sous-catégorisant pas obligatoirement un objet direct.

En s'appuyant sur un travail expérimental en psycholinguistique, Stallings *et al.* (1998) font des observations similaires concernant le HNPS. Les auteurs ont mis en place trois expériences pour tester l'hypothèse selon laquelle un verbe fréquemment utilisé dans une construction *shiftée* (construction dans laquelle l'objet n'est pas adjacent au verbe) a tendance à favoriser le HNPS. Plus précisément, Stallings *et al.*

s'intéressent aux verbes pour lesquels l'objet direct peut être réalisé comme une proposition subordonnée. Ces auteurs montrent que, lorsque les locuteurs doivent produire une phrase contenant un verbe, son objet direct sous forme de SN et un SP complément ou modifieur, ils produisent significativement plus de constructions à HNPS (V SP SN) avec les verbes pouvant réaliser leur objet direct sous la forme d'une subordonnée, qu'avec les autres verbes. Cela signifie que les locuteurs ont tendance à produire plus de HNPS avec des verbes acceptant par ailleurs un complément phrastique. En plus des données sur corpus relevées par Wasow (1997), les données expérimentales en production renforcent l'idée selon laquelle certaines classes syntaxiques de verbe favorisent une construction spécifique.

Wasow (1997) propose d'expliquer ces observations en s'appuyant sur le traitement *online* de l'énoncé par le locuteur. Il estime que, pour faciliter la planification et la production d'un énoncé, un locuteur a intérêt à s'engager le plus tard possible sur le type de complémentation qu'il va choisir. Ainsi, lorsque l'objet direct peut être réalisé sous la forme d'une complétive, le locuteur a intérêt à produire l'objet direct en dernier, afin de se laisser un maximum de temps pour décider si l'objet va être réalisé comme une complétive ou comme un SN. De même, en ce qui concerne les verbes qui sous-catégorisent optionnellement un objet direct, en produisant l'objet direct le plus tard possible, le locuteur conserve le plus longtemps possible l'opportunité d'utiliser la version transitive indirecte du verbe. Selon Wasow (1997), c'est donc la flexibilité autorisée par des propriétés syntaxiques spécifiques à certains verbes qui est mise à profit par le locuteur, afin de faciliter au maximum la production de son énoncé.

L'interprétation de Stallings *et al.* (1998) repose sur l'idée que les locuteurs sont sensibles à la fréquence d'apparition des verbes dans une construction spécifique. Dans la mesure où une complétive est un constituant souvent lourd et complexe, les verbes pouvant prendre une complétive sont très facilement séparés de leur objet direct par un SP (26-a) ou un adverbe (26-b) modifieur.

(26) a. *Mary **said** in a loud voice [that Bill would sing]*
b. *Mary **learned** yesterday [that she would be allowed to go hiking]*

Les auteurs estiment que la fréquence d'apparition d'un verbe dans ce type de construction influence significativement la production de HNPS pour ce verbe. Ils émettent une hypothèse qu'ils nomment *Verb disposition hypothesis* et qu'ils formulent de la façon suivante : « *chaque verbe porte en lui des informations sur l'histoire de sa participation dans des structures shiftées et cette histoire influence la probabilité qu'il autorise le HNPS* » [15].

La nature des cadres de sous-catégorisation du verbe semble donc avoir une influence sur l'ordre des compléments en anglais. D'autres propriétés du lemme verbal jouent un rôle, notamment le lien sémantique unissant le verbe et l'un de ses dépendants.

15. « *individual verbs carry with them information on the history of their participation in shifted structures and [...] this history influences the likelihood of their allowing HNPS* » (Stallings *et al.*, 1998, p. 396).

5.5.3. Lien sémantique entre le verbe et un constituant

Lorsqu'il existe un lien sémantique entre le verbe et l'un des constituants, l'ordre d'apparition des éléments postverbaux s'en trouve affecté. En effet, un constituant fortement lié sémantiquement au verbe va avoir tendance à apparaître adjacent à la tête verbale. Cette idée avait déjà été émise par Behaghel (1932) :

> « *La loi suprême est celle selon laquelle ce qui est étroitement lié mentalement, se positionne aussi à proximité* » [16].

Dans l'exemple attesté (27), le SN objet est séparé du verbe par le modifieur *as closely as possible*.

(27) *…replicate as closely as possible the Brown Corpus…* (Wasow & Arnold, 2003, p. 11)

Le modifieur est plus long, plus complexe, et l'objet n'est pas nouveau d'un point de vue informationnel car cette phrase est précédée d'un paragraphe sur le corpus Brown. Wasow (2002) explique que, en tant que modifieur très générique, l'interprétation de *as closely as possible* est très sensible au contexte et, dans le cas présent, c'est le verbe qui va permettre de l'interpréter. Ainsi, avec le modifieur strictement adjacent au verbe, l'organisation de la phrase reflète directement la dépendance sémantique à travers l'ordre des mots. Ce qui motive cet ordre, c'est donc la relation sémantique entre *replicate* et *as closely as possible*. Trois études de corpus mettent en lumière l'effet du lien sémantique entre le verbe et l'un des constituants sur l'ordre des syntagmes postverbaux.

Premièrement, Wasow (2002) propose une étude à partir de 827 SV contenant une des cinq paires verbe-préposition suivantes : *attribute … to*, *bring… to*, *obtain… from*, *share… with*, *take… into*. À l'image de ce qui a été présenté dans la section 5.4, il a annoté ces SV selon trois classes exclusives relatives à la notion de collocation :

1. pas de collocation (***share*** *that cost* ***with others***),
2. collocation sémantiquement transparente (***bring*** *the debate* ***to an end***),
3. collocation sémantiquement opaque, idiome (***take*** *our concerns* ***into account***).

Dans le cas des collocations opaques (idiomes), l'interprétation du verbe et du SP dépend de la co-occurrence des deux éléments, tandis que, dans le cas des collocations transparentes, il n'y a pas de dépendance sémantique, chaque élément pouvant être interprété indépendamment de l'autre. Le lien sémantique entre le verbe et le SP est donc plus étroit dans les idiomes que dans les collocations transparentes. Si l'on recoupe les 3 classes d'idiomacité avec l'ordre des constituants (pas de HNPS (V SN SP) vs. HNPS (V SP SN)), on obtient les résultats suivants :

1. pas de collocation, 15% de HNPS

16. « *Das oberste Gesetz ist dieses, dass das geistig eng Zusammengehörige auch eng zusammengestellt wird* » (Behaghel, 1932)

2. collocation transparente, 47% de HNPS
3. collocation opaque, 60% de HNPS

Les différences de proportions sont statistiquement significatives. Ainsi, la différence entre les deux types de collocations (2 et 3), située au niveau de la transparence sémantique, se retrouve dans la différence de proportions de structure HNPS. La corrélation entre le lien sémantique et l'ordre des constituants dans le phénomène du HNPS est donc bien réelle.

Deuxièmement, Hawkins (2000) s'intéresse à l'ordre relatif des SP postverbaux et montre l'importance de la dépendance sémantique entre le verbe et les SP dans le choix de l'ordre. Pour cela, il applique un test d'implication (*entailment test*) permettant de déterminer si le verbe est sémantiquement dépendant du SP et réciproquement. En ce qui concerne la dépendance du verbe par rapport au SP, le test consiste à vérifier que le sens du SV contenant le SP implique le sens du SV sans le SP. Si tel est le cas, le verbe est indépendant du SP ; dans le cas contraire, le verbe est dépendant du SP. De cette façon, on peut contraster les SP selon le lien qu'ils entretiennent avec le verbe dans la phrase *Pat lived in Paris for about a year* : *lived* est indépendant de *for about a year*, car *Pat lived in Paris for about a year* implique *Pat lived in Paris* ; tandis que *lived* est dépendant de *in Paris*, car *Pat lived in Paris for about a year* n'implique pas *Pat lived for about a year*. Hawkins applique le même test pour évaluer la dépendance du SP au verbe : le SV contenant le verbe a-t-il le même sens que le SV contenant une expression verbale sémantiquement minimale, telle que *do something* ? Ainsi, dans la phrase *Pat talked with Chris about Sandy*, le SP *with Chris* est sémantiquement indépendant du verbe, car *Pat talked with Chris about Sandy* implique *Pat did something with Chris* ; alors que le SP *about Sandy* est dépendant du verbe, car *Pat talked with Chris about Sandy* n'implique pas *Pat did something about Sandy*. En s'appuyant sur ces tests, Hawkins a annoté plusieurs centaines de phrases selon les dépendances du verbe et des SP. Il en déduit que le verbe et le SP ont tendance à être adjacents quand il y a une dépendance sémantique.

Troisièmement, Wasow & Arnold (2003) travaillent sur la construction verbe-particule en utilisant le même test que Hawkins (2000) pour déterminer si la particule est dépendante du verbe et réciproquement. Ils observent que la particule a tendance à apparaître adjacente au verbe, lorsqu'elle est sémantiquement dépendante de ce dernier. Il faut moduler cette conclusion en fonction de la contrainte de poids : le poids du SN objet reste le facteur déterminant dans un grand nombre de cas, et ce n'est qu'une fois ce facteur sous contrôle que l'on peut observer les effets du lien sémantique entre verbe et particule.

Les trois études s'appuient sur deux actualisations de la notion de lien sémantique entre deux constituants : la première repose sur le classement des liens sémantiques sur l'échelle de l'idiomaticité, la seconde sur un test d'implication cherchant à déterminer, entre deux constituants, une dépendance sémantique orientée. Ces deux approches vont dans le même sens et montrent que, en anglais, le lien sémantique entre verbe et constituant joue un rôle sur l'ordre dans les phénomènes de HNPS, construction verbe-particule et ordre relatif des SP multiples.

5.5.4. Pour le français

À notre connaissance, le caractère animé des référents n'a jamais été pointé comme un facteur pouvant influencer l'ordonnancement des constituants postverbaux en français. Il serait intéressant d'étudier son impact dans la perspective de l'hypothèse *directe* relative à l'influence du caractère animé sur l'ordre des constituants. Étant donné que le phénomène étudié ne met en jeu que la linéarisation des compléments, il constitue un lieu d'observation de l'effet direct du caractère animé sur l'ordre des mots, indépendamment de l'assignation des fonctions grammaticales.

Les travaux sur le français ne font pas directement référence à l'influence des rôles sémantiques. Néanmoins, Schmitt (1987a,b) pointe l'importance de la sémantique du verbe dans le choix de l'ordre des compléments. Les verbes concernés sont des verbes de mouvement impliquant un point de départ ou un état initial et un « *but à atteindre* » qui peut être le « *point final d'une évolution, d'une action ou d'une réflexion* ». L'auteur ajoute les verbes exprimant l'union ou la division, pour lesquels « *à partir de nombreux éléments d'origine se forme une unité* » ou « *la subdivision de cette unité présente […] le but du devenir* » (Schmitt, 1987a, p. 294). Enfin, il intègre les verbes impliquant une relation de comparaison entre les deux compléments, avec « *le point de départ […] toujours situé à côté du verbe* » et « *le complément se rapportant au résultat ou figurant comme solution […] à la fin de la phrase* » (Schmitt, 1987a, p. 295). Il cite les verbes suivants :

- *passer de* SN *à* SN,
- *faire de* SN SN ,
- *préférer* SN *à* SN
- *remplacer* SN *par* SN,
- *troquer* SN *contre* SN,
- *traduire de* SN *à* SN,
- *transposer de* SN *à* SN,
- *joindre* SN *à* SN,
- *lier* SN *à* SN,
- *séparer* SN *de* SN,
- *confondre* SN *avec* SN,
- *substituer* SN *par* SN

Il affirme que, pour ces verbes, « *l'ordre des objets dans la partie postverbale de la phrase […] [est] prédéterminé par le verbe ou plus précisément, ses caractéristiques sémantiques, indépendamment du contexte et de la qualité du* CO^1 *et du* CO^2 *dans la phrase* » (Schmitt, 1987a, p. 290). Ses conclusions reposent sur l'observation de données recueillies manuellement dans des oeuvres de Jean-Paul Sartre et de Marguerite Duras, ainsi que dans des articles du *Monde* et du *Nouvel Observateur*. Ne trouvant aucune occurrence de l'ordre inverse à celui présenté ci-dessus, Schmitt affirme que la plupart des verbes cités imposent un ordre à leurs compléments.

À la suite de Schmitt (1987a,b), nous émettrons l'hypothèse selon laquelle la sémantique du verbe influence l'ordre de ses compléments. Cependant, nous considérons que la sémantique du verbe ne constitue qu'une contrainte préférentielle. Elle n'impose pas mais favorise plus ou moins fortement un ordre plutôt que l'autre.

5.6. Hiérarchies relatives au discours

Nous revenons en détail sur les hiérarchies concernant le caractère défini et l'opposition information nouvelle/information donnée. Nous évoquerons également les effets possibles de la familiarité et de la référentialité.

5.6.1. Caractère défini

La nature du déterminant d'un SN est un élément formel qui semble avoir une influence dans l'ordonnancement des arguments du verbe selon la hiérarchie suivante :

(28) défini $\prec$ indéfini

Concernant l'alternance dative en anglais, Ransom (1979) a montré que le caractère défini du thème et du destinataire ont une influence sur l'acceptabilité de la construction à double objet (V SN SN). Les observations sur corpus faites par Collins (1995), Gries (2003b) et Bresnan *et al.* (2007) révèlent qu'un destinataire avec un déterminant défini favorise la construction à double objet, alors qu'un thème défini privilégie la construction à SP datif.

Au sujet de l'allemand, Bader & Häussler (2010) constatent que, dans les propositions subordonnées, l'ordre entre le sujet (SUJ) et l'objet indirect au datif (OI) est influencé par le caractère défini des deux constituants. Lorsque les deux constituants sont définis, l'ordre OI - SUJ apparaît dans 44% des cas, tandis que, lorsque le SUJ est indéfini et l'OI défini, la proportion d'ordre OI-SUJ atteint 83%. Ainsi, l'ordre relatif du SUJ et de l'OI respecte la tendance exprimée par la hiérarchie (28).

5.6.2. Information nouvelle - information donnée

Les arguments du verbe s'organisent selon l'accessibilité des informations qu'ils portent. La tendance générale peut se résumer de la façon suivante : « *exprimer ce qui est donné avant ce qui est nouveau par rapport à cela* »[17] (Gundel, 1988). La hiérarchie en (29) exprime la même idée.

(29) information donnée $\prec$ information nouvelle

Cette hiérarchie pose un problème de définition, dans la mesure où la notion d'information donnée ou nouvelle varie selon les auteurs. Les travaux dont nous rendrons compte dans la suite de cette section s'appuient sur la typologie de Prince (1981) et notamment sur la distinction *donné dans le discours*, *inférable* et *nouveau dans le discours*.

À partir de données extraites du corpus *Aligned-Hansard*, Arnold *et al.* (2000) ont observé l'effet du statut des référents (*donné* vs. *nouveau*) dans les phénomènes du HNPS et de l'alternance dative en anglais. Après annotation selon les trois catégories de Prince (1981), Arnold *et al.* ont intégré les *inférables* aux *donnés*, étant donné que

17. « *State what is given before what is new in relation to it* » (Gundel, 1988, 229)

la catégorie *inférable* était très peu représentée. Concernant le HNPS, leurs données contiennent 390 occurrences de constructions *bring (…) to* et *take (…) into account*. Lorsque le SN et le SP sont de même longueur, les auteurs observent plus de 30% de HNPS (V SP SN) avec un SN nouveau, alors qu'il y en a moins de 5% avec un SN donné. De même, lorsque le SN est un peu plus long que le SP (un à trois mots de plus), la proportion de HNPS atteint près de 55% si le SN est nouveau, tandis qu'elle se situe autour de 7% si le SN est donné. Ces données suivent la tendance décrite dans la hiérarchie (29), selon laquelle un objet direct *nouveau* tend à favoriser le HNPS : V SP $SN_{nouveau}$. En ce qui concerne l'alternance dative, Arnold *et al.* ont relevé 269 occurrences du verbe *give* accompagné de deux compléments réalisés soit sous la forme de la construction à double objet (V $SN_{destinataire}$ SN_{theme}), soit sous la forme de la construction à SP datif (V SN_{theme} $SP_{destinataire}$). Lorsque le thème et le destinataire sont de longueur égale, les auteurs constatent que la construction à double objet s'observe à un peu plus de 15% pour un destinataire *nouveau* et un thème *donné*, alors que cette même construction est attestée à près de 60% pour un destinataire *donné* et un thème *nouveau*. Le choix de la construction semble, une fois de plus, affecté par le statut du référent dans le discours.

Certains travaux (Ariel, 1990; Arnold, 1998) ont montré que le poids et le statut informationnel sont hautement corrélés. Dans une phrase, les éléments récemment mentionnés tendent à avoir besoin d'une description moins complexe, du fait qu'ils sont accessibles pour le locuteur et l'interlocuteur, tandis que les éléments nouveaux tendent à être exprimés à l'aide de constituants plus complexes. Cette forte corrélation amène à questionner l'intérêt d'utiliser les deux notions pour rendre compte de l'ordre des compléments du verbe. Cette question a notamment été soulevée par Hawkins (1994). Ce dernier estime que la seule notion de poids suffit à expliquer les données. Arnold *et al.* (2000) apportent des arguments réfutant ce point de vue, en s'appuyant sur des données de corpus, ainsi que sur une expérience d'élicitation. Ils affirment que « *la complexité grammaticale et le statut dans le discours influencent l'ordre des constituants* » [18]. Ces auteurs montrent notamment que malgré l'existence d'une importante corrélation, les deux facteurs sont significatifs lorsque l'ordre des constituants est modélisé à partir des données de corpus. L'expérience d'élicitation tend à confirmer les observations sur corpus. La modélisation de l'ordre choisi par les sujets est significativement améliorée par les deux facteurs, poids et statut des référents.

À l'image de ce qui a été vu sur les hiérarchies de poids, on peut expliquer la hiérarchie (29) en termes de traitement, c'est-à-dire du point de vue de l'interlocuteur, et en termes de planification et de production, c'est-à-dire du point de vue du locuteur. D'une part, reprendre ce qui a été dit précédemment permet de créer de la continuité dans le discours et de faciliter ainsi la compréhension de l'interlocuteur (Arnold *et al.*, 2000, p. 32). D'autre part, les constituants contenant de l'information nouvelle sont plus difficiles à produire pour au moins trois raisons. Premièrement, accéder à l'item

18. « *both grammatical complexity and discourse status influence constituent order* » (Arnold *et al.*, 2000, p. 51)

lexical ou choisir l'expression linguistique appropriée prend plus de temps. Deuxièmement, planifier la phonétique est plus difficile si l'expression n'a pas déjà été amorcée (*primed*). Troisièmement, au niveau articulatoire, le locuteur peut être sujet à plus de difficultés lorsque les items lexicaux n'ont pas été produits précédemment (Bock, 1986, 1987; Bock & Irwin, 1980). De plus, faire référence à des constituants 'donnés' est moins compliqué, dans la mesure où les représentations conceptuelles et linguistiques sont déjà activées (Arnold *et al.*, 2000, p. 33). Les entités sont donc plus facilement accessibles (Bock & Warren, 1985; Prat-Sala & Branigan, 2000). Ainsi, produire d'abord les constituants faisant référence à ce qui a été dit et, ensuite, les constituants comportant l'information nouvelle, permet d'optimiser le temps de planification et de production du locuteur dans la parole spontanée : le locuteur produit d'abord les éléments donnés, ce qui lui laisse du temps pour résoudre les difficultés liées aux éléments nouveaux.

5.6.3. Familiarité

La hiérarchie faisant référence à la familiarité est reproduite en (30).

(30) thème plus familier $\prec$ thème moins familier $\prec$ commentaire [19]

Il nous semble que cette hiérarchie, utilisée par Allan (1987) et Siewierska (1993), regroupe au moins deux dimensions : d'une part, la familiarité du locuteur avec un référent ou avec le thème du discours, et d'autre part, l'organisation informationnelle d'un énoncé (thème vs. commentaire, par exemple).

La familiarité du locuteur avec le référent dont il parle peut influer sur l'accessibilité de ce référent et donc sur son ordre d'apparition lors de la linéarisation. Ce principe a été mentionné pour expliquer l'ordre des conjoints dans les coordinations, par exemple chez Cooper & Ross (1975) (*me first principle*) et chez Allan (1987). Ainsi, dans la coordination de deux noms propres désignant un couple, le premier nom à apparaître a tendance à être celui référant à la personne la plus proche du locuteur. Une notion similaire apparaît dans le travail de Ertel (1977) visant à dégager les facteurs influençant la sélection du sujet syntaxique en allemand. À partir d'expériences psycholinguistiques, Ertel montre que, en plus de l'agentivité, la proximité (*closeness*) entre le référent du sujet et l'ego du locuteur a une influence sur le choix du sujet. La hiérarchie de familiarité implique une dimension personnelle et émotive. Cette dernière dépend de la subjectivité de chaque locuteur. La tendance générale consiste à placer les éléments les plus familiers en premier. Cela implique que ce qui apparaît effectivement en premier peut dépendre du locuteur produisant la séquence.

Allan (1987) et Siewierska (1993) font le lien entre l'idée de référent familier et la notion de thème (*topic*), car un élément qui n'est pas familier tend à être ancré dans un thème familier. Ainsi, les phrases ont tendance à s'organiser selon le schéma : *ce dont ont parle* $\prec$ *ce qu'on en dit*. Les notions de thème et de commentaire renvoient,

19. 'Thème' est la traduction de l'anglais *topic*, et 'commentaire' de *comment*.

chez Allan (1987) et Siewierska (1993), à l'organisation du discours au-delà de la phrase, comme en témoignent les exemples présentés par ces auteurs.

5.6.4. Pour le français

Blinkenberg mentionne l'idée que, pour la langue écrite, si le complément d'objet indirect « *ne fait que répéter ou rappeler un terme connu* » (Blinkenberg, 1928, p. 181), alors ce dernier peut apparaître naturellement avant le complément d'objet direct. Pour sa part, Berrendonner (1987) estime que la tendance à rencontrer l'ordre '*information donnée* $\prec$ *information nouvelle*' est la conséquence d'un phénomène plus large : la dernière position dans l'ordre linéaire est le lieu privilégié du focus. Cet auteur définit le focus comme « *la partie [du contenu de l'énoncé] qui est présentée […] comme l'information maximalement pertinente au regard de l'état courant du savoir partagé* » (Berrendonner, 1987, p. 10). Il considère que le statut de focus et le statut d'information nouvelle vont souvent de pair, ce qui explique la présence de l'information nouvelle en dernière position. Les trois principaux arguments avancés par l'auteur pour justifier l'idée selon laquelle l'ordre des mots sert à marquer la focalisation sont de trois ordres. Premièrement, selon l'auteur et ses informateurs, il est plus naturel de placer en dernière position le contenu répondant à une question partielle, comme dans les paires (31) et (32) (Berrendonner, 1987, p. 10).

(31) a. *À qui Moscou a-t-il envoyé un message d'appui ?*
b. *Moscou a envoyé un message d'appui au mouvement des cent-un.*

(32) a. *Qu'est-ce que Moscou a envoyé au mouvement des cent-un ?*
b. *Moscou a envoyé au mouvement des cent-un un message d'appui*

Deuxièmement, Berrendonner estime que seul l'élément placé en dernière position peut être coordonné à un focus contrastif, comme dans l'exemple (33) (Berrendonner, 1987, p. 11).

(33) *Solon donna donc au peuple les droits civils,* et non les droits politiques

Troisièmement, l'auteur constate une « *nette tendance à […] disposer [les compléments] dans un ordre canonique qui va du défini à l'indéfini* ». L'élément en dernière position a donc tendance à être accompagné d'un déterminant indéfini, ce qui, selon Berrendonner, est cohérent avec son statut de focus.

Les arguments apportés par Berrendonner (1987) ne nous semblent pas convaincants, dans la mesure où la paire question-réponse en (34) [20] ainsi que la phrase

20. Ajoutons que, dans leur étude sur la relation entre prosodie et structure informationnelle en français, Beyssade *et al.* (2004b) exemplifient leur propos à l'aide de paires 'question/réponse' où le SN focalisé n'apparaît pas en dernière position, prouvant que cet ordre est tout à fait acceptable :

(i) a. *Qu'est-ce que Jean-Pierre a offert à ton fils ?*
b. *Jean-Pierre a offert un train électrique à mon fils*

(ii) a. *Qu'est-ce que tu as donné aux étudiants de Licence pour le concours blanc ?*
b. *J'ai donné trois exercices de syntaxe aux étudiants de Licence pour le concours blanc*

(35) sont tout à fait acceptables et ne sont pas moins naturelles que les exemples de l'auteur.

(34) a. *Qu'est-ce que Moscou a envoyé au mouvement des cent-un ?*
b. *Moscou a envoyé un message d'appui au mouvement des cent-un.*

(35) *Solon donna donc les droits civils au peuple,* et non les droits politiques

De plus, la tendance selon laquelle les constituants définis précèdent les indéfinis ne repose sur aucune donnée autre que l'intuition de l'auteur. L'argument ne semble donc pas avoir de poids. *Contra* Berrendonner (1987), nous considérons donc que la dernière position n'est pas la place privilégiée accueillant le constituant focalisé.

Nous reprenons la caractérisation du focus étroit, envisagé dans l'articulation du contenu de l'énoncé en 'fond/focus', donnée par Beyssade *et al.* (2004a) pour le français : « *le SX qui est explicitement asserté ou questionné est réalisé comme un syntagme autonome portant sur son bord droit un ton lié à la force illocutoire de l'énoncé* » [21]. Le focus étroit se caractérise donc par le présence d'un ton spécifique sur le bord droit du constituant. L'étude de la relation entre l'ordre des compléments postverbaux et la réalisation du focus doit impérativement prendre en compte la dimension prosodique de l'énoncé. Dans le cadre de ce travail, nous n'avons pas conduit d'étude acoustique qui pourrait permettre de statuer sur la relation entre focus et ordre des mots. Étant donné que nous laissons de côté la notion de focus, les deux hypothèses à étudier concernent le statut *donné* ou *nouveau* des référents ainsi que le caractère défini ou non du déterminant introduisant les constituants.

Les travaux sur le français présentent des contraintes en lien avec celles rencontrées dans la littérature sur les autres langues. Cependant, le problème de l'ordonnancement des compléments postverbaux en français n'a pas fait l'objet d'étude sur corpus ou de travaux expérimentaux. Seul Schmitt (1987a,b) a travaillé sur un large échantillon de phrases attestées, ce qui lui a permis de mettre en lumière un aspect jusque là ignoré, à savoir l'influence de la sémantique du verbe sur l'ordre de ses compléments. Dans la mesure où la problématique de l'ordre des compléments verbaux ne met en jeu que des contraintes préférentielles, il apparaît indispensable de mener une étude quantitative sur corpus et/ou des expériences psycholinguistiques. C'est ce que nous nous proposons de faire dans le chapitre suivant. Néanmoins, nous ne pourrons pas étudier la totalité des contraintes abordées dans ce chapitre. Le travail que nous allons présenter constitue une étude exploratoire qui ne rend pas compte de façon exhaustive du problème de l'ordre des constituants postverbaux. Il s'agira principalement d'étudier des contraintes relatives aux trois grandes hiérarchies que nous avons présentées et de statuer sur leur influence en français.

---

21. « *the XP that is specifically asserted or questioned, is realized as an autonomous phrase bearing on its right edge a tone related to the illocutionary force of the utterance* » (Beyssade *et al.*, 2004a, p. 472)

Chapitre 6

Analyse de données de corpus

Sommaire

L'objectif de ce chapitre est d'explorer la question de l'ordonnancement des compléments du verbe en français, à travers une étude sur des données de corpus. Nous cherchons à identifier les contraintes générales agissant sur l'ordre des constituants qui apparaissent en position postverbale et appelés par la tête verbale. De ce fait, nous prenons en compte des données variées, parmi lesquelles :

– des verbes sous-catégorisant un SP datif

(1) *[…]* ***donner*** *un emploi à chacun* (FTB)

– des verbes sous-catégorisant un SP locatif

(2) *[…]* ***trouver*** *une agence de voyage dans le quartier* (ESTER)

– des constructions verbales spécifiques

(3) *[…]* ***faire*** *du tourisme une véritable industrie* (FTB)

– des verbes à construction support

(4) *[…]* ***faire*** *la traduction à ses coéquipiers* (Est-Républicain)

– des expressions figées non-connexes

(5) *[…]* ***Mettre*** *la Sept en difficulté* (FTB)

Nous nous attacherons à proposer une formalisation des contraintes mises en jeu, en utilisant les techniques décrites dans le chapitre 2 et déjà mises en oeuvre dans la partie concernant la position de l'adjectif.

L'étude de l'ordre des compléments postverbaux soulève un problème méthodologique de recueil de données. Extraire les données appropriées nécessite, dans l'idéal, un corpus annoté syntaxiquement dans lequel les fonctions des dépendants du verbe sont mentionnées. Il existe un seul corpus de ce type en français : le *French Treebank* (Abeillé & Barrier, 2004). Or, dans ce corpus, le nombre de données pertinentes pour le problème qui nous intéresse est relativement faible. Il a donc été nécessaire de compléter les données issues du *French Treebank* avec celles d'autres corpus. Ces derniers ne présentant pas l'annotation nécessaire à une extraction automatique et intégrale, il a fallu procéder à des extractions partielles en utilisant une méthode d'échantillonnage. Dans un premier temps, nous avons réalisé une table de données sur la base d'une méthode d'échantillonnage qui s'est révélée trop rudimentaire lors de l'analyse des données, notamment en raison de l'importance du lemme verbal dans le phénomène étudié. Nous avons alors procédé à un deuxième échantillonnage, moins naïf, qui a abouti à une nouvelle table de données.

Le présent chapitre s'organise en trois grandes parties. Dans la première, nous présenterons l'étude préliminaire qui a permis de mettre à jour le problème d'échantillonnage. À partir de la table de données préliminaire, nous comparerons les différentes mesures de poids envisageables. Ensuite, nous exposerons les éléments permettant d'affirmer que le lemme verbal influe sur le choix de l'ordre de ses com-

pléments. Dans la partie suivante, nous présenterons la table de données qui est au centre de ce travail. Après avoir décrit la manière dont la table a été obtenue et annotée, nous exposerons les variables étudiées, ainsi que les premières observations que nous pouvons en faire. Ensuite, la modélisation du phénomène sera présentée. Dans la troisième partie, nous reviendrons en détail sur trois aspects particulièrement importants d'après nos résultats : le lemme verbal, le caractère animé des référents et leur statut *donné* ou *nouveau*.

6.1. Étude préliminaire

6.1.1. Méthode

Dans la mesure où notre objet d'étude est l'ordre relatif des compléments post-verbaux, nous avons choisi de retenir les cas où le verbe est suivi uniquement de deux constituants sous-catégorisés. Les deux patrons qui nous intéressent sont donc **V SX SP** et **V SP SX** où SX est l'objet direct du verbe réalisé soit comme un SN (6-a), soit comme une subordonnée (6-b), soit comme une infinitive (6-c), et le SP est sous-catégorisé par le verbe.

(6) a. *[…] donna* ***des idées*** *aux gérants de sicav court terme* (FTB)
b. *[…] a dit à Christophe* ***qu'un type lui cherchait des noises*** (Est-Républicain)
c. *[…] demande à Gbagbo* ***de refuser*** (ESTER)

Afin de constituer la table de données préliminaire, nous avons extrait ces deux patrons des trois corpus suivants : *French Treebank* (FTB), Est-Républiacain (ER) et ESTER [1].

6.1.1.1. Extraction de FTB

Dans un premier temps, en nous appuyant sur l'annotation syntaxique et fonctionnelle du FTB, nous avons récupéré automatiquement l'ensemble des verbes ditransitifs accompagnés de deux compléments sous-catégorisés. Plus exactement, nous avons extrait chaque verbe suivi d'un constituant ayant la fonction OBJ ainsi que d'un SP ayant la fonction A_OBJ, DE_OBJ ou P_OBJ. De plus, à la droite du verbe, le noeud dominant le noyau verbal ne devait pas contenir d'autres constituants que les deux sous-catégorisés [2]. Un exemple de patron extrait du FTB est présenté dans la figure 6.1 : le verbe *donne* est suivi du SP, *à ses concitoyens*, ayant la fonction A_OBJ et du SN objet *l'impression d'avoir perdu toute capacité d'initiative*. Nous avons ainsi extrait 409 occurrences des patrons contenant 159 lemmes verbaux différents.

1. Pour une présentation de ces trois corpus, voir la section 2.1.3 du chapitre2.
2. Dans le FTB, le noyau verbal regroupe le verbe, les clitiques et les auxiliaires sous la forme d'une structure plate dominée par un noeud VN (noyau verbal). Les propositions finies ne présentent pas de SV. Le verbe fini et ses compléments forment une structure plate au même niveau que le

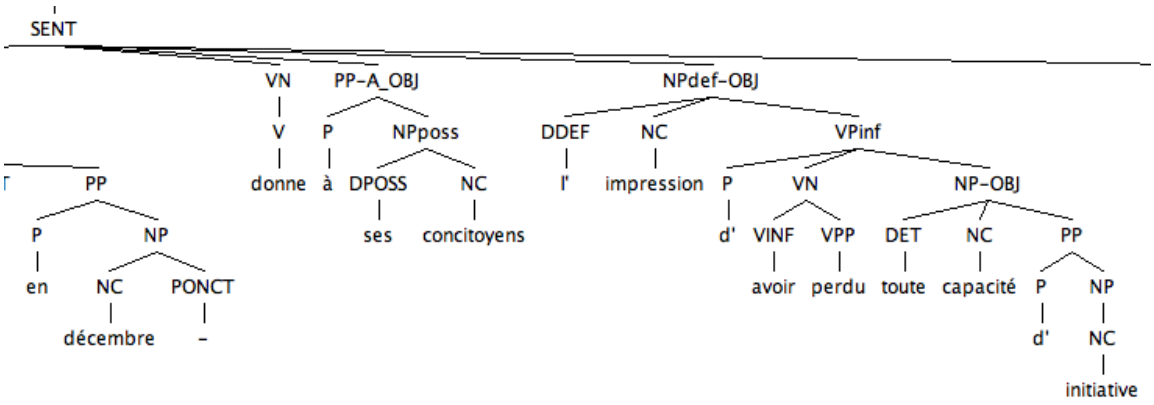

FIGURE 6.1.: Patron V SP SN extrait du FTB et visualisé à l'aide de l'interface graphique de Tregex (Levy & Andrew, 2006).

6.1.1.2. Extraction de ER et ESTER

L'extraction des données de ER et de ESTER n'a pas pu se faire automatiquement, dans la mesure où ces deux corpus ne comportent qu'une annotation en parties du discours. Il a donc fallu choisir une méthode d'extraction et de sélection des données. Nous avons procédé à une sélection manuelle des phrases à partir des lemmes verbaux : après avoir extrait toutes les occurrences d'un verbe potentiellement ditransitif, nous avons choisi les contextes respectant les deux patrons recherchés. Étant donné la taille des corpus, il n'était pas possible de passer en revue l'intégralité des occurrences de ces verbes. Nous avons donc procédé à un échantillonnage au niveau des lemmes verbaux. Pour cela, nous avons listé les verbes ditransitifs du FTB, ainsi que leur fréquence brute dans ce même corpus. Nous avons sélectionné les 18 lemmes les plus fréquents et nous les avons cherchés dans ER [3]. Le nombre de patrons retenus pour chaque lemme est proportionnel à sa fréquence dans FTB. Dans le cas de ESTER, certains lemmes retenus n'apparaissaient pas dans les contextes recherchés et, pour d'autres, le nombre d'occurrences était relativement faible. Nous avons donc élargi le nombre de lemmes recherchés. De ce corpus, nous avons extrait 24 verbes [4], dont 12 en commun avec ceux de ER. *In fine*, nous avons sélectionné 577 phrases pour ER et 307 pour ESTER. Ces 884 phrases ont fait l'objet d'une analyse syntaxique automatique, suivie d'une correction manuelle.

sujet du verbe.

3. La liste exacte de ces lemmes est : *ajouter*, *annoncer*, *assurer*, *devoir*, *dire*, *donner*, *expliquer*, *faire*, *mettre*, *montrer*, *passer*, *permettre*, *porter*, *prendre*, *réduire*, *rendre*, *trouver* et *vendre*.

4. La liste exacte est : *accorder*, *ajouter*, *annoncer*, *appeler*, *assurer*, *demander*, *dire*, *donner*, *expliquer*, *faire*, *lancer*, *mettre*, *montrer*, *obtenir*, *offrir*, *passer*, *permettre*, *porter*, *prendre*, *présenter*, *proposer*, *réduire*, *rendre* et *trouver*.

6.1.1.3. Description de la table préliminaire

Le table est composée de 1293 phrases annotées syntaxiquement. Les lemmes sont au nombre de 161 : les 159 du FTB, auxquels s'ajoutent deux lemmes présents dans ER et ESTER, mais pas dans FTB (*dire* et *montrer*) [5]. Les données ont été organisées sous forme d'une table que nous nommons *Table Préliminaire*, abrégée sous la forme *TP*. Chaque ligne de cette table contient les informations relatives à une phrase du corpus. Elle comprend notamment l'ordre des constituants postverbaux, encodé sous la forme d'une variable binaire :

`ordre`

= 0 : l'ordre attesté est SX_{OBJ} SP,
= 1 : l'ordre attesté est SP SX_{OBJ}.

La table *TP* comprend également le lemme verbal, le corpus d'origine de la phrase et la forme sous laquelle est réalisé l'objet direct (`realObjet` = 1, quand l'objet direct est un SN, `realObjet` = 0 lorsqu'il s'agit d'une subordonnée ou d'une infinitive). Un extrait de *TP* est présenté dans la table 6.1.

ordre	lemVb	realObjet	corpus
1	permettre	0	ER
1	obtenir	1	FTB
0	justifier	1	FTB
0	réserver	1	FTB

Table 6.1.: Extrait de la table *TP*

La table *TP* contient 1293 occurrences de verbes accompagnés de deux compléments et répartis en 161 lemmes verbaux. On observe 573 phrases avec l'ordre SX_{OBJ} SP (44.3%) et 720 avec l'ordre SP SX_{OBJ} (55.7%). Cependant, cette première observation doit être modulée en fonction du corpus et de la nature de l'objet, comme le montre la table 6.2.

Premièrement, si l'on distingue les cas où l'objet est réalisé comme SN des autres cas, la tendance s'inverse : pour `realObjet` = SN, l'ordre SX_{OBJ} SP est représenté à 57.7%. Cela s'explique par le fait que, lorsque l'objet est réalisé sous la forme d'une subordonnée, il est quasi systématiquement postposé au SP [6]. De plus, dans le cas d'un objet nominal, il y a d'importantes variations selon les corpus : 67.8% d'ordre SN SP pour le FTB, 63.6% pour ESTER et 46.3% pour ER. Cette observation pourrait suggérer que le type de discours a une influence sur l'ordre choisi.

5. Ces deux verbes apparaissent avec deux compléments dans le FTB. Toutefois, aucune de ces occurrences ne correspond aux deux patrons recherchés, car les deux compléments ne sont pas les seuls constituants présents à la droite du verbe.

6. L'unique occurrence de subordonnée apparaissant avant le SP est une interrogative indirecte qui fait partie d'une expression figée :

(i) *montrer [de quel bois il se chauffait] [à son rival]* (corpus ER)

	realObjet = SN				realObjet ≠ SN			
	ER	ESTER	FTB	*TP*	ER	ESTER	FTB	*TP*
ordre = 0	193	150	229	572	1	0	0	1
	46.3%	63.6%	67.8%	57.7%	0.6%	0%	0%	0.3%
ordre = 1	224	86	109	419	159	71	71	301
	53.7%	36.4%	32.3%	42.2%	99.4	100%	100%	99.7%
Totaux	417	236	338	991	160	71	71	302
	100%	100%	100%	100%	100%	100%	100%	100%

TABLE 6.2.: La variable ordre en fonction de corpus et de realObjet

6.1.2. Analyse

Nous allons étudier deux aspects de notre problématique à partir de la table préliminaire. Dans un premier temps, nous reviendrons sur la notion de poids afin d'en estimer l'effet et d'en proposer une mesure. Dans un deuxième temps, nous montrerons que le lemme verbal a une importante influence sur l'ordonnancement de ses compléments.

6.1.2.1. Le poids

Afin d'observer l'effet du poids sur l'ordre des compléments postverbaux, nous utilisons deux mesures de longueur et deux mesures de complexité syntaxique : la longueur des constituants en nombre de mots et en nombre de syllabes, la complexité des constituants en nombre de noeuds syntaxiques et en nombre de noeuds syntagmatiques [7]. Cette démarche vise à déterminer quelle mesure rend le mieux compte de l'ordre observé.

La longueur en mots Elle a été calculée à partir de l'annotation syntaxique des phrases. Par exemple, le nombre de mots d'un SN correspond au nombre de feuilles du sous-arbre dominé par le noeud SN. Nous avons encodé la longueur à l'aide de trois variables longSXobjMots, longSPmots et longRelMots.

longSXobjMots : nombre de mots du syntagme ayant la fonction objet (SN, subordonnée ou infinitive) ;

longSPmots : nombre de mots du SP ;

7. Nous réduisons ici la notion de complexité syntaxique au nombre de noeuds, c'est-à-dire à une mesure de la profondeur de la structure. L'objectif de notre travail n'est pas de définir la notion de complexité. Ainsi, les mesures de poids que nous utilisons ne prennent pas en compte d'autres paramètres qui semblent intervenir dans la notion de complexité en psycholinguistique : accessibilité des référents (Gibson, 2000) ou activation d'un mot (Vasishth, 2003). Le modèle de complexité présenté dans Blache (2010) permet, en intégrant ces divers paramètres, d'obtenir une quantification de la notion psycholinguistique de complexité. Notons que d'autres contraintes préférentielles, que nous développerons dans la suite de ce chapitre, prennent en compte une partie des paramètres de la complexité, notamment le statut du référent (donné/nouveau) et le caractère défini.

longRelMots : longueur relative : longSXobjMots− longSPmots.

Les moyennes de longSXobjMots et longSPmots sont : μ(longSXobjMots) = 7.14, μ(longSPmots) = 5.02. Ces moyennes montrent que les SX_{OBJ} ont tendance à être plus longs que les SP. Cela s'explique notamment par la présence des subordonnées et des infinitives qui, à elles seules, ont une moyenne de 11.67 mots. Les graphiques de la figure 6.2 donnent une représentation de la relation entre longueur en mots et ordre des compléments postverbaux.

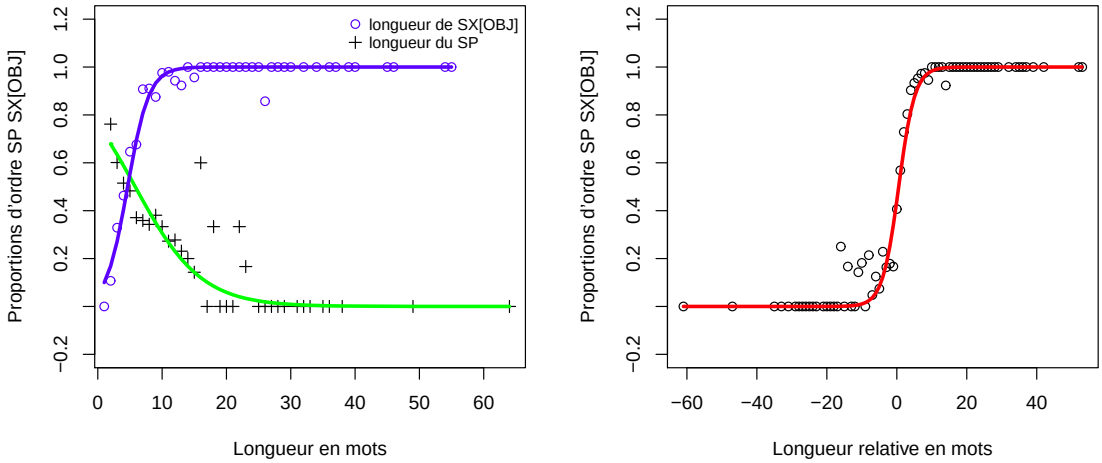

FIGURE 6.2.: À gauche, longSXobjMots et longSPmots en fonction de ordre avec les courbes logistiques résumant le mieux les données de *TP* ; à droite, longRelMots en fonction de ordre et la courbe logistique résumant le mieux les données de *TP* .

Dans le graphique de gauche, on observe clairement que lorsque la longueur du SP augmente, la proportion d'ordre SP SX_{OBJ} diminue (courbe verte). Inversement, plus la longueur du SX_{OBJ} est importante, plus l'ordre SP SX_{OBJ} est fréquent (courbe bleue). Le graphique de droite montre la relation entre longueur relative et proportion d'ordre SP SX_{OBJ}. La variable longRelMots prend une valeur négative lorsque le SP est plus long que le SX_{OBJ} et une valeur positive lorsque c'est le SX_{OBJ} qui est plus long que le SP. Quand longRelMots est compris entre -20 et 15, on observe une variation de l'ordre, avec une tendance claire : plus le SX_{OBJ} est long par rapport au SP, plus on observe d'occurrences de l'ordre SP SX_{OBJ}. Au-delà de ces deux valeurs pour longRelMots (longRelMots $<$ -20 ou longRelMots $>$ 15), il n'y a pas de variation d'ordre dans les données. L'ajustement des données à la courbe de régression indique que la longueur relative en nombre de mots est, à elle seule, un excellent prédicteur de l'ordre des compléments postverbaux.

La longueur en syllabes Nous disposons de la longueur en syllabes pour la sous-partie du corpus extraite du FTB [8]. Nous avons relevé la longueur du syntagme objet (`longSXobjSyll`) et celle du SP (`longSPsyll`).

Afin de comparer l'influence de la longueur en syllabes et en mots sur l'ordre relatif des compléments verbaux, nous présentons les deux graphiques de la figure 6.3, qui contiennent la proportion d'ordre SP SX_{OBJ} dans la sous-partie extraite du FTB, en fonction de la longueur en syllabes (graphique de gauche) et de la longueur en mots (graphique de droite).

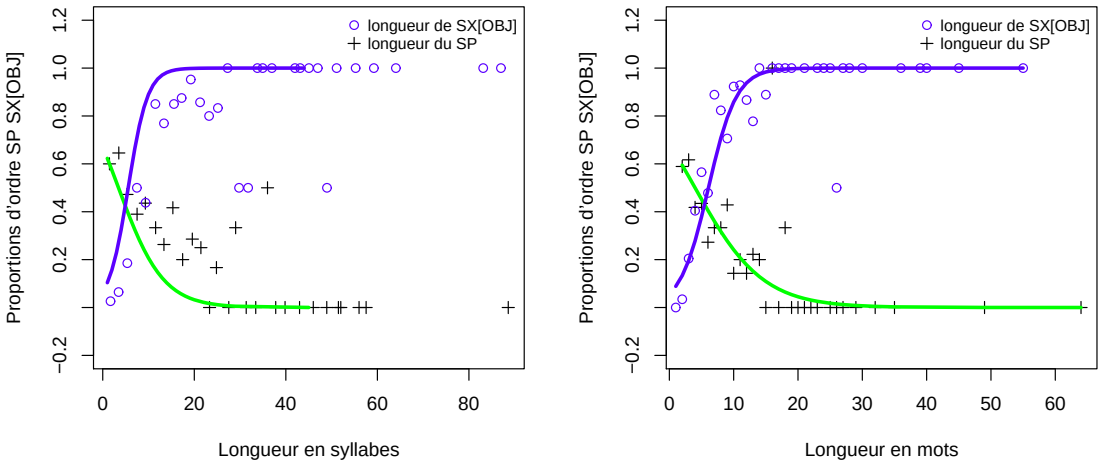

FIGURE 6.3.: À gauche, `longSXobjSyll` et `longSPsyll` en fonction de `ordre`, avec les courbes logistiques résumant le mieux les données de la sous-partie du corpus extraite de FTB ; à droite, `longSXobjMots` et `longSPmots` en fonction de `ordre` et les courbes logistiques résumant le mieux les données de *TP* .

On observe que les courbes de régression présentent une allure générale similaire, ce qui indique que l'effet de ces deux mesures de longueur est convergent. Par rapport à la droite de régression, les données relatives à la longueur syllabique semblent légèrement plus dispersées que celles concernant la longueur en mots.

Nous avons calculé le pourcentage de données se conformant au principe *court avant long* en fonction des deux mesures. D'après le nombre de syllabes, 78.5% sont conformes à ce principe ; d'après la longueur en mots, on obtient un taux de 86.5% [9].

8. La syllabation du corpus a été effectuée à l'aide du logiciel ELITE. Pour plus de détails, voir la section 4.2.1 du chapitre 4.

9. Ces pourcentages ne tiennent pas compte des phrases dans lesquelles le nombre de syllabes ou de mots est le même pour les deux constituants, dans la mesure où le principe *court avant long* ne peut pas s'appliquer.

Pour rendre compte de l'ordre relatif des compléments du verbe, la mesure de longueur la plus adaptée est donc le nombre de mots.

Le nombre de syllabes ne dit pas grand chose sur la complexité des constituants, notamment dans le cas où ces derniers sont relativement courts, puisqu'un mot de trois ou quatre syllabes peut augmenter la taille d'un constituant sans que ce dernier soit complexe. La longueur en mots donne une meilleure approximation de la complexité, dans la mesure où l'ajout de mots va souvent de pair avec une complexité syntaxique plus importante. Ce résultat semble indiquer que ce n'est pas la longueur qui est centrale dans le choix de l'ordre. Ce dernier dépend plutôt d'une mesure davantage liée à la complexité.

La complexité en noeuds syntaxiques Pour un SP donné, nous avons compté le nombre de noeuds non-terminaux dominés par l'étiquette SP. Cette mesure est dépendante du schéma d'annotation. Celui du FTB est relativement plat et ne contient aucune catégorie vide. Le nombre de noeuds syntaxiques est donc lié au nombre de mots. Par exemple, pour le SN *le directeur*, le nombre de noeuds syntaxiques est 3 : le noeud SN, le noeud Nom et le noeud Déterminant. Comme pour la longueur en mots, nous disposons de trois variables encodant la complexité : `longSXobjNds`, `longSPnds` et `longRelNds`.

`longSXobjNds` : nombre de noeuds syntaxiques non-terminaux contenus dans le syntagme ayant la fonction objet (SN, subordonnée ou infinitive) ;

`longSPnds` : nombre de noeuds syntaxiques non-terminaux contenus dans le SP ;

`longRelNds` : complexité relative : `longSXobjNds` − `longSPnds`.

De même que pour la longueur, les moyennes des nombres de noeuds syntaxiques montrent que les SP tendent à être plus courts que les SX_{OBJ} : μ(`longSXobjNds`) = 12.28 et μ(`longSPnds`) = 8.96. Nous avons représenté l'impact des trois variables sur l'ordre des compléments verbaux à l'aide de deux graphiques, présentés dans la figure 6.4.

Ces représentations permettent d'observer les effets attendus : plus le SX_{OBJ} contient de noeuds syntaxiques, plus la proportion d'ordre SP SX_{OBJ} augmente, et inversement, plus le SP contient de noeuds syntaxiques, plus la proportion d'ordre SP SX_{OBJ} diminue. Le graphique de droite représente l'influence des deux mesures de complexité combinées. Comme pour le nombre de mots, les données sont bien ajustées à la courbe de régression. La variable `longRelNds` est donc un bon prédicteur de l'ordre des compléments postverbaux. Nous reviendrons à la fin de cette section sur la comparaison des mesures de poids.

Complexité en noeuds syntagmatiques Afin d'affiner la mesure de la complexité, nous avons évalué le nombre de constituants syntaxiques composant le SX_{OBJ} et le SP. Pour cela, nous avons compté le nombre de noeuds syntagmatiques [10] conte-

10. Les noeuds syntagmatiques du FTB sont : "Srel" (proposition relative), "Sint" (proposition conjuguée interne), "Ssub" (proposition subordonnée), "VPinf" (proposition infinitive), "VPpart"

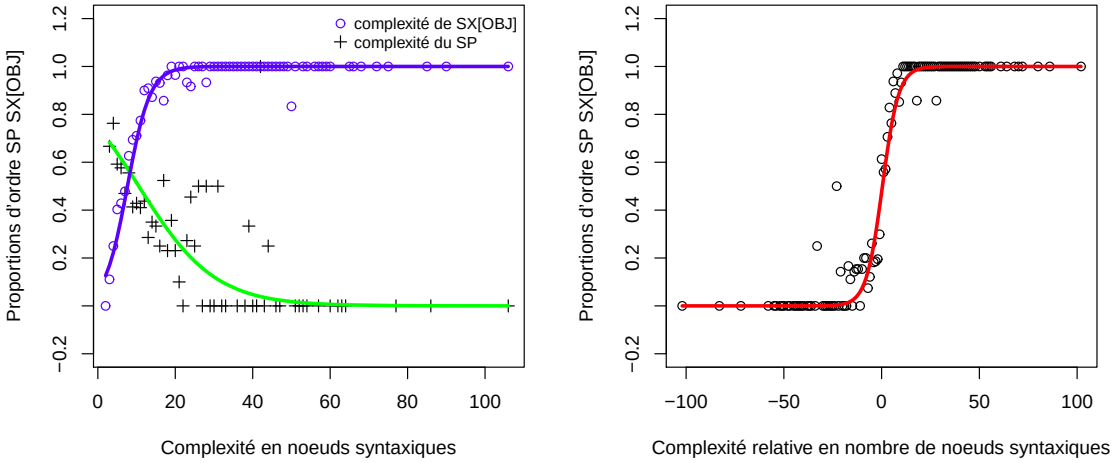

FIGURE 6.4.: À gauche, `longSXobjNds` et `longSPnds` en fonction de `ordre` avec les courbes logistiques résumant le mieux les données de *TP* ; à droite, `longRelNds` en fonction de `ordre` et la courbe logistique résumant le mieux les données de *TP*.

nus dans chacun de ces deux syntagmes et nous avons établi trois variables : `longSXobjSynt`, `longSPsynt` et `longRelSynt`.

`longSXobjSynt` : nombre de noeuds syntagmatiques contenus dans le syntagme ayant la fonction objet (SN, subordonnée ou infinitive) ;

`longSPsynt` : nombre de noeuds syntagmatiques contenus dans le SP ;

`longRelSynt` : complexité relative : `longSXobjSynt` − `longSPsynt`.

Nous représentons à nouveau le comportement de ces trois variables dans deux graphiques (figure 6.5). Les observations sont similaires à celles que nous avons proposées pour le nombre de noeuds syntaxiques.

Les graphiques permettent d'observer que les trois mesures − nombre de mots, nombre de noeuds syntaxiques et nombre de noeuds syntagmatiques − sont pertinentes pour évaluer le poids des constituants en relation avec la problématique de l'ordre des compléments. Nous comparons ces trois mesures en caractérisant plus précisément leur pouvoir prédictif sur la variable `ordre`. Pour cela, nous calculons le pourcentage de données obéissant au principe *court avant long*, comme nous l'avons fait pour départager la longueur en mots et en syllabes. Les résultats sont présentés dans la table 6.3.

Les proportions de données respectant le principe de poids croissant montrent que les trois mesures sont quasi équivalentes du point de vue du phénomène qui

(proposition participiale), "COORD" (syntagme coordonné), "NP" (syntagme nominal), "PP" (syntagme prépositionnel), "AP" (syntagme adjectival), "AdP" (syntagme adverbial).

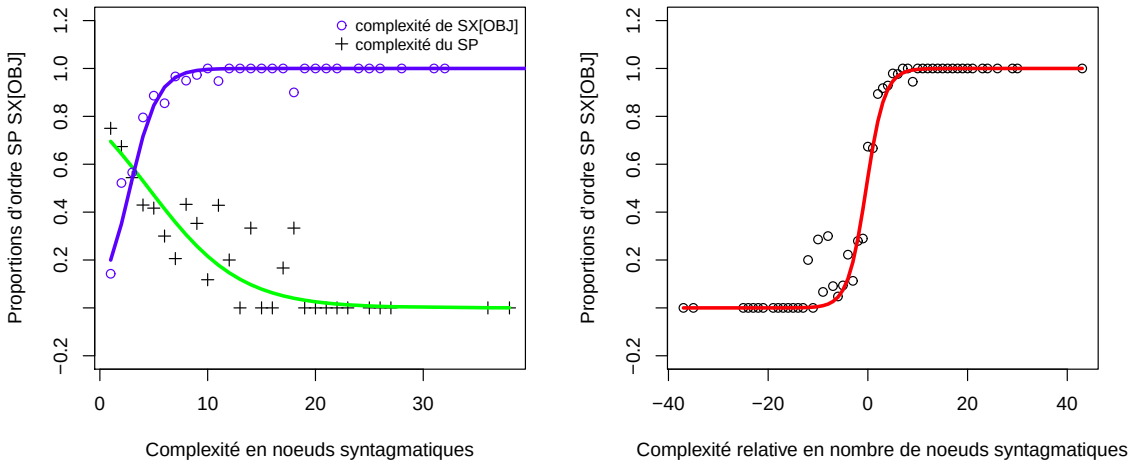

FIGURE 6.5.: À gauche, `longSXobjSynt` et `longSPsynt` en fonction de `ordre` avec les courbes logistiques résumant le mieux les données de *TP* ; à droite, `longRelSynt` en fonction de `ordre` et la courbe logistique résumant le mieux les données de *TP*.

Mots	Noeuds syntaxiques	Noeuds syntagmatiques
86.7%	85.4%	83.6%

TABLE 6.3.: Pourcentage des données de *TP* se conformant au principe *court avant long* selon les trois mesures de poids.

nous intéresse. Ces observations sont très similaires à celles que Wasow (1997) a faites sur l'anglais. Aucune des trois mesures ne semble clairement se détacher comme le meilleur prédicteur. Étant donné que la mesure en nombre de mots est la plus facile à obtenir, nous l'utiliserons dorénavant pour approximer le poids des constituants. Retenons, cependant, que la mesure de longueur sans référence à la composition des constituants (longueur en syllabes) s'avère être un moins bon prédicteur pour l'ordre des compléments postverbaux. Cela indique que la complexité syntaxique est un élément plus important que la longueur *stricto sensu*.

Dans la section 6.1.1.3, nous avons remarqué que lorsque l'objet est réalisé sous la forme d'une subordonnée ou d'une infinitive, l'ordre attesté est à plus de 99% SP SX_{OBJ}. Cela signifie que la variation est quasiment nulle dans ces données. Il semble donc intéressant d'observer l'effet du poids dans la sous-partie de la table de *TP* qui présente une plus grande flexibilité dans l'ordonnancement des compléments. Nous appelons *TPbis* la table contenant les données pour lesquelles l'objet est un SN, autrement dit pour lesquelles `realObjet` = 1. Tel que cela est montré dans la table

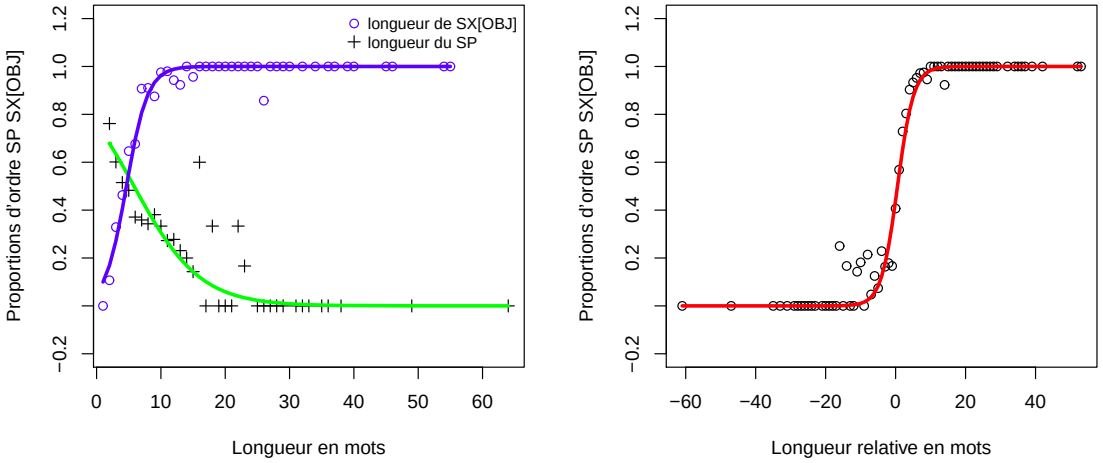

FIGURE 6.6.: À gauche, `longSXobjMots` et `longSPmots` en fonction de `ordre` avec les courbes logistiques résumant le mieux les données de *TPbis* ; à droite, `longRelMots` en fonction de `ordre` et la courbe logistique résumant le mieux les données de *TPbis*.

6.2, *TPbis* contient 991 phrases. L'effet du poids est relativement similaire dans *TP* et dans *TPbis*, comme en témoignent les graphiques de la figure 6.6. De plus, la part de données de *TPbis* qui obéit au principe *court avant long* est de 85.5%.

6.1.2.2. Le verbe

L'identité du verbe introduisant les deux compléments a une influence sur l'ordre attesté. Pour étayer cette affirmation, nous comparons le comportement des douze verbes les plus fréquents de *TP*.

Les données de la table 6.4 indiquent que les verbes *permettre*, *demander*, *annoncer*, *dire*, *faire*, *prendre* et *mettre* ont une préférence pour l'ordre SP SN, tandis que les verbes *trouver*, *porter*, *donner* et *passer* favorisent l'ordre SN SP. Les préférences de chaque lemme s'observent à divers degrés : alors que *mettre* ne présente qu'une légère préférence pour l'ordre SP SN, les verbes *permettre* et *demander* favorisent largement cet ordre. Notons que les verbes qui présentent la préférence la plus forte pour l'ordre SP SN sont des verbes sous-catégorisant des subordonnées et des infinitives.

Ces premières observations sont confirmées lorsque l'on prend en compte la longueur. Le graphique de gauche de la figure 6.7 représente le comportement de la variable `ordre` en fonction de la longueur [11] pour trois des verbes les plus fréquents de *TP* : *faire*, *montrer* et *donner*. Ce graphique permet d'observer que la longueur

11. Dans ces graphiques, la longueur est exprimée sur une échelle logarithmique afin de réduire la dispersion des données et de rendre l'effet du lemme verbal plus visible.

	ordre = 0		ordre = 1		Totaux	
permettre	1	0.6%	172	99.4%	173	100%
demander	6	14%	37	86%	43	100%
annoncer	6	19.4%	25	80.6%	31	100%
dire	6	20%	24	80%	30	100%
faire	28	27.7%	73	72.3%	101	100%
montrer	29	28.2%	74	71.8%	114	100%
prendre	16	43.2%	21	56.8%	37	100%
mettre	58	46.8%	66	53.2%	124	100%
trouver	20	69%	9	31%	29	100%
porter	26	74.3%	9	25.7%	35	100%
donner	71	78%	20	22%	91	100%
passer	39	97.5%	1	2.5%	40	100%

TABLE 6.4.: Comportement de la variable `ordre` en fonction des 12 verbes les plus fréquents de *TP*.

influe sur l'ordre choisi pour les trois verbes, mais que chaque verbe conserve un comportement différent une fois la longueur prise en compte.

L'influence du verbe sur la variable `ordre` présente également des différences si l'on prend en compte la préposition introduisant le SP sous catégorisé. Nous définissons une nouvelle variable, `lemPrep`, qui correspond à la concaténation du lemme verbal et de la préposition réprésentée à l'aide de trois valeurs : *à*, *de* et *autres*. Cette variable donne une approximation des divers emplois que peuvent avoir les lemmes verbaux. Dans la table 6.5, nous indiquons les valeurs de `lemPrep` pour les verbes *faire*, *prendre* et *mettre*. Pour les verbes, *prendre* et *faire*, on observe que le type de préposition marque une très grande différence dans le choix de l'ordre, tandis que pour le verbe *mettre*, la préposition ne semble pas capter des différences de préférence. Les données de la table *TP* indiquent que, pour certains verbes, en plus de l'identité du lemme, le type d'usage qui en est fait peut avoir une influence sur l'ordonnancement de ses compléments.

lemPrep	ordre = 0		ordre = 1	
faire à	25	96.2%	1	3.8%
faire de	3	4%	72	96%
mettre à	5	55.6%	4	44.4%
mettre autres	53	46.1%	62	53.9%
prendre à	7	100%	0	0%
prendre autres	9	30%	21	70%

TABLE 6.5.: La variable `lemPrep` en fonction de la variable `ordre`.

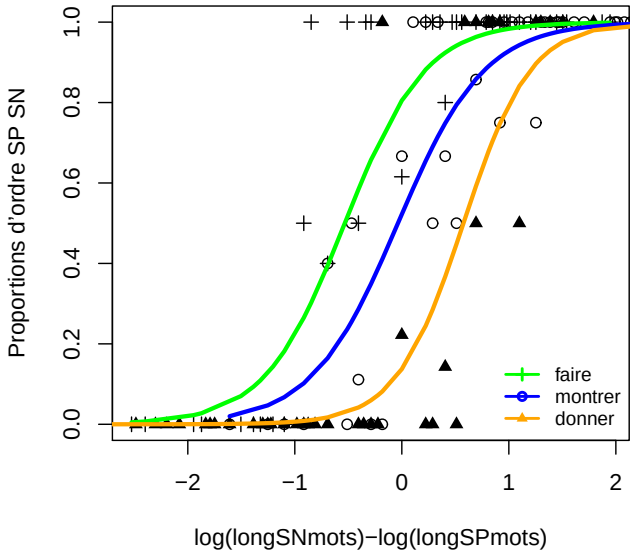

FIGURE 6.7.: Longueur relative des constituants (échelle logarithmique) en fonction de `ordre` pour les verbes *faire*, *montrer* et *donner*, avec les courbes logistiques résumant les données relatives à chaque verbe.

Étant donné que la variable `ordre` présente de très importantes variations selon l'identité du lemme verbal et son emploi, la méthode d'échantillonnage adoptée pour l'élaboration de la table préliminaire ne semble pas la mieux adaptée. En effet, les données issues de ER et ESTER ne contiennent qu'une vingtaine de verbes. Afin de mieux rendre compte du phénomène de l'ordonnancement des compléments verbaux, il est nécessaire de mieux représenter la diversité des lemmes verbaux. De plus, nous avons noté qu'il existe une véritable alternance d'ordre lorsque l'objet du verbe est réalisé sous la forme d'un SN (`realObjet` = 1). Il apparaît que l'enjeu de notre problématique se situe au niveau de la description et de la formalisation de l'ordre relatif du SN et du SP. Pour la suite de ce travail, il est donc nécessaire de réduire le phénomène étudié au cas où l'objet du verbe est un SN.

La table de données étudiée dans la section suivante respecte les deux enseignements tirés de l'étude préliminaire : l'échantillonnage respecte la diversité des verbes et leur proportion dans le FTB ; les données ne contiennent que des réalisations nominales de l'objet.

6.2. Étude de la table de données finale

Dans cette section, nous proposons l'analyse de la table finale, que nous désignerons dorénavant sous le nom de *TF*. En plus des données extraites du FTB, de ER et de

ESTER, cette table contient du matériel provenant du corpus C-ORAL-ROM, corpus d'oral spontané [12].

6.2.1. Méthode

6.2.1.1. Extraction des données

La table *TF* est construite à partir de l'extraction des patrons **V SN SP** et **V SP SN**, où le SN et le SP sont sous-catégorisés par le verbe. Les données extraites de FTB sont les mêmes que celles de la table *TP*, à savoir 338 phrases contenant 159 lemmes verbaux différents. En ce qui concerne les trois autres corpus, nous avons sélectionné manuellement les phrases contenant les patrons à partir des lemmes verbaux. Cependant, à la différence de ce que nous avons fait pour *TP*, nous avons élargi au maximum le nombre de lemmes différents extraits dans chaque corpus. De plus, pour échantillonner les verbes, nous ne nous sommes pas appuyée sur la fréquence brute dans le FTB, mais sur la fréquence du verbe en emploi ditransitif avec une préposition spécifique. Nous avons procédé de cette façon afin d'approximer la proportion de chaque emploi d'un même lemme (cf. les différences observées selon les valeurs de `lemPrep`, dans la section 6.1.2.2). Ainsi, nous avons sélectionné 150 lemmes verbaux dans le corpus ER, avec un nombre d'occurrences allant de 1 à 37. Le nombre total de phrases extraites de ce corpus est de 782. Pour ESTER, nous avons trouvé les deux patrons pour 65 lemmes verbaux qui se répartissent dans 204 phrases [13]. Enfin, nous avons extrait 110 phrases contenant les deux patrons du corpus CORAL. Ces 110 occurrences représentent 42 lemmes verbaux. Notons que le nombre d'occurrences de ESTER et de CORAL n'est pas toujours proportionnel à celui trouvé dans FTB. En effet, dans la mesure où les deux patrons sont plus difficiles à trouver dans les corpus oraux, nous avons fait le choix de ne pas limiter le nombre d'occurrences par lemme.

6.2.1.2. Description de la table finale

La table *TF* contient 1434 occurrences des patrons, qui se répartissent en 182 lemmes verbaux. L'ordre SN SP se rencontre pour 1010 occurrences, soit 70.4%. L'ordre SP SN représente 424 occurrences, soit 29.6%.

La table 6.6 présente les proportions observées dans chaque corpus. Ainsi, contrairement à ce qui a été observé pour *TP* (cf. table 6.2), l'ordre SN SP est sous représenté dans le FTB, par rapport aux autres corpus. De plus, les différences inter-corpus

12. Ce corpus est présenté dans la section 2.1.3 du chapitre 2.

13. Pour *TF* nous avons restreint nos recherches à une sous-partie du corpus ESTER, composée des transcriptions de France Inter, France Info, France Culture et Radio Classique. Pour la table *TP*, nous avions utilisé l'intégralité du corpus ESTER qui, en plus des transcriptions que nous venons d'énumérer, contient des transcriptions d'émissions diffusées sur RFI et sur la Radio Télévision Marocaine. Nous avons fait ce choix pour *TF* dans le but de réduire la variation qui pourrait exister entre le français métropolitain et le français parlé en dehors de France.

(maximum de moins de 10 points entre FTB et ESTER) sont moindres par rapport à celles de la table *TP* (environ 20 points d'écart entre FTB et ER). Étant donné que l'échantillonnage de *TF* est plus respectueux de la distribution du FTB, on peut émettre l'hypothèse que les proportions de *TF* sont plus représentatives des préférences pour l'ordre SN SP dans chaque corpus.

	CORAL		ER		ESTER		FTB		Totaux	
`ordre` = 0	81	73.6%	544	69.6%	156	76.5%	229	67.8%	1010	70.4%
`ordre` = 1	29	26.4%	238	30.4%	48	23.5%	109	32.2%	424	29.6%
	110	100%	782	100%	204	100%	338	100%	1434	100%

TABLE 6.6.: La variable `ordre` en fonction de `corpus`, dans *TF*

6.2.1.3. L'annotation du corpus

Afin d'étudier le rôle du caractère animé et celui de la sémantique du verbe sur l'ordre des compléments, nous avons procédé à l'annotation manuelle des 1434 phrases de *TF*. Nous avons utilisé les outils d'annotation suivants : *Multi-purpose Annotation Environment* (MAE) et *Multi-document Adjudication Interface* (MAI) (Stubbs, 2011). Le premier a servi à l'annotation et le second au processus d'unification des jugements (adjudication).

Le caractère animé Le caractère animé est une propriété inhérente des référents. On considère généralement qu'il s'agit d'une propriété hiérarchique allant de l'humain à l'inanimé. Pour l'annotation, nous avons repris la hiérarchie utilisée par Zaenen *et al.* (2004) et présentée dans la table 6.7. La description des catégories est détaillée dans le guide d'annotation, en grande partie inspiré de celui de Garretson (2004) et reproduit dans l'annexe B. L'objectif était d'annoter, en contexte, les référents des SN objets des verbes et des SN objets des prépositions dans les SP. Nous avons donc annoté les 2868 référents.

HUMAIN >	ANIMÉ >	INANIMÉ
humain	animal	concret
	organisation	non-concret
	machine intelligente	lieu
	véhicule	temps

TABLE 6.7.: Hiérarchie du caractère animé.

L'annotation a été réalisée par trois annotateurs : l'un d'entre eux était l'auteur de ce texte, les deux autres étaient des étudiants en Licence 3 de linguistique informatique à l'université Paris Diderot [14]. L'annotation s'est faite en deux étapes.

14. Ces deux annotateurs ont été rémunérés pour ce travail, dans le cadre d'un stage financé par l'UMRi 001 Alpage (Paris Diderot - INRIA).

Dans un premier temps, les trois annotateurs ont travaillé de manière indépendante à l'aide du guide d'annotation, en attribuant à chacun des 2868 référents une des neufs étiquettes présentées en 6.7. Dans un deuxième temps, les trois annotateurs ont examiné ensemble les cas de désaccord et ont choisi une étiquette définitive sur la base du consensus. À la fin du processus d'annotation, nous disposons de trois corpus annotés par trois personnes et d'un corpus de référence issu de la mise en commun des trois annotations et de la prise de décision commune sur les désaccords.

Nous évaluons la qualité de l'annotation à l'aide de deux mesures d'accord. Nous choisissons la terminologie de Artstein & Poesio (2008) qui, à notre connaissance est l'article le plus récent faisant le point sur ces mesures. Nous estimons la fiabilité des annotations avec le Multi-π de Fleiss, aussi connu sous le nom de κ de Carletta (Carletta, 1996), et le Multi-κ, aussi appelé généralisation de Cohen (Cohen, 1960). Il est admis que, au-delà de 0.8, ces mesures témoignent d'un bon accord inter-annotateur et donc de données fiables. Pour l'annotation du caractère animé des référents, nous obtenons : Multi-π= 0.83 et Multi-κ = 0.87 (k=3, N=2868). De plus, nous avons réduit le caractère animé à deux catégories en regroupant, sous l'étiquette *animé*, les catégories *humain*, *animal*, *organisation* ; et, sous l'étiquette *inanimé*, les autres catégories. Pour cette catégorisation sémantique binaire, on obtient des mesures d'accord révélant que nos données sont très fiables : Multi-π = 0.91 et Multi-κ = 0.93 (k=3, N=2868). Afin de donner une idée de la distribution des catégories et des désaccords, la table 6.8 représente la superposition des trois matrices de confusion pour les trois paires d'annotateurs. Pour chacune de ces paires, les étiquettes des lignes correspondent à un annotateur et les étiquettes des colonnes à l'autre annotateur. La fréquence des accords se trouve donc dans la diagonale du tableau et la fréquence des désaccords dans les autres cases.

On observe que la plupart des désaccords se concentrent autour de l'étiquette *non-concret*, notamment pour les paires *concret/non-concret*, *lieu/non-concret*, *organisation/non-concret* et *humain/non-concret*. Il existe également de nombreux désaccords pour la paire *humain/organisation*. Ces lieux de désaccords sont similaires à ceux rencontrés par Zaenen *et al.* (2004) et témoignent souvent d'une différence d'interprétation en contexte.

Dans le cadre du travail sur l'ordre des compléments postverbaux, nous utiliserons la distinction *animé* vs. *inanimé*. La table 6.9 représente la fréquence des accords et des désaccords relatifs à cette distinction pour les trois paires d'annotateurs.

Nous illustrons les cas de désaccords qui touchent à la distinction *animé/inanimé*. Premièrement, nous observons que la limite entre *organisation* et *non-concret* est souvent difficile à tracer, comme dans les exemples (7)-(8). Pour les SP de ces deux phrases, deux annotateurs ont choisi l'étiquette *organisation*, tandis que le troisième a opté pour *non-concret*.

	Ani	Conc	Hum	Lieu	N-conc	Orga	Temps	Véh	Mac	Oanim
Ani	46	4	2	0	0	0	0	0	0	4
Conc		684	4	110	196	10	0	20	6	4
Hum			1521	3	118	109	0	1	0	4
Lieu				286	120	4	0	0	0	1
N-conc					4518	207	60	11	8	18
Orga						429	0	0	0	6
Temps							46	0	0	0
Véh								14	2	0
Mac									0	0
Oanim										0

TABLE 6.8.: Matrice de confusion pour l'annotation du caractère animé de *TF* (*Oanim* signifie 'je ne sais pas'.)

	Animé	Inanimé
Animé	2121	400
Inanimé		6062

TABLE 6.9.: Matrice de confusion pour l'annotation de l'opposition *animé* vs *inanimé*.

(7) *le gouvernement de M. Pierre Bérégovoy et M. Gomez [...]* ***cèdent*** *[l'usine Eisswein et l'électroménager de Thomson SA]$_{SN}$ [à un groupe familial étranger, l'italien Elettro Finanziaria Spa]$_{SP}$* (FTB)

(8) *il* ***fait*** *[du groupe français]$_{SP}$ [le numéro un mondial en équipements de transmissions]$_{SN}$* (FTB)

Dans la version de référence, l'étiquette *organisation* a été sélectionnée pour *un groupe familial* (7), considérant que le bénéficiaire de la transaction est la ou les personnes à la tête du groupe. À l'inverse, nous avons choisi l'étiquette *non-concret* dans le cas de *groupe français* (8), puisque l'objet subissant la transformation est une entité abstraite 'entreprise' plutôt qu'un groupe de personnes.

Deuxièmement, certains contextes posent des problèmes pour différencier les *humains* du *non-concret*. C'est le cas du SN *26 personnes* dans l'exemple (9), pour lequel deux annotateurs ont utilisé l'étiquette *non-concret*, alors que le troisième a choisi *humain*. En ce qui concerne les *emprunteurs* en (10), le SN peut renvoyer soit à des individus qui empruntent, soit à des sociétés ou des banques. Pour ce référent, l'étiquette *humain* a été utilisée deux fois, et l'étiquette *non-concret* une fois.

(9) *Ce qui devrait* ***porter*** *[notre entreprise]$_{SN}$ [à 26 personnes]$_{SP}$* (ER)

(10) *Sinon, si la banque n'a pas dans sa clientèle immédiate un débiteur idéal, elle* ***élargira*** *[le champ de ses investigations]$_{SN}$ [aux emprunteurs qui ont fait savoir qu'ils sont prêts, le cas échéant, à émettre des transactions privées]$_{SP}$.* (FTB)

Dans la version de référence du corpus, nous avons choisi la catégorie *non-concret* pour la phrase (9), puisque le SN fait référence à une quantité plus qu'à des êtres humains. Pour la phrase (10), nous avons opté pour l'étiquette *humain*, considérant qu'il s'agissait bien d'individus, et non pas de sociétés dans un sens abstrait.

Ces exemples illustrent la complexité de la catégorisation sémantique en contexte. Étant donné la difficulté de la tâche, les mesures d'accord obtenues apparaissent très satisfaisantes. Les données relatives au caractère animé utilisées dans la suite du chapitre sont celles du corpus de référence.

La sémantique du verbe La sémantique du verbe présente un double intérêt pour l'étude de l'ordre des compléments du verbe. D'une part, elle permet de désambiguïser différents emplois d'un même verbe en contexte. D'autre part, on peut émettre l'hypothèse que l'ordre des compléments est influencé par la sémantique et que le classement des verbes permettra de faire émerger des tendances.

La classification sémantique des verbes étant un travail allant bien au-delà du cadre de notre thèse, nous avons utilisé une ressource existante, disponible et, à notre connaissance, la plus étendue : le dictionnaire *Les Verbes du Français* (LVF) de Dubois & Dubois-Charlier (1997). Il s'agit d'une ressource écrite à la main qui contient 25 610 entrées verbales représentant 12 310 lemmes verbaux classés selon leurs propriétés syntactico-sémantiques. La classification s'appuie sur l'analyse des types de sujets, de compléments et d'adjoints (animé, inanimé, abstrait, singulier/pluriel, collectif…), sur le type de réalisation des arguments (SN, SP, proposition…), ainsi que sur les alternances syntaxiques autorisées par le verbe. Pour l'annotation, nous utilisons trois des cinq niveaux de classification disponibles :

niveau 5 : classes génériques (codé par une lettre majuscule)

niveau 4 : classes sémantico-syntaxiques (codé par un chiffre)

niveau 3 : sous-classes syntaxiques (codé par une lettre minuscule)

Nous illustrons les différents niveaux de classification avec un exemple concret (11).

(11) *[...] elle* ***cède*** *à celui-ci 3,5% de la SGAB et 19,6% de la ACESA [...]* (FTB)

La classification du verbe *céder* en contexte est *D2a* :

niveau 5 : classe générique **D** (don, privation)

niveau 4 : classe sémantico-syntaxique D**2** (donner qq ch à qq'un, obtenir qq ch de qq'un)

niveau 3 : sous-classe syntaxique D2**a** (fournir qq ch à qq'un)

Classes génériques Elles sont au nombre de quatorze. Elles indiquent le sens général du verbe. La liste exacte est la suivante :

C : communication
D : don, privation
E : entrée, sortie
F : frapper, toucher
H : états physiques et comportements
L : locatif
M : mouvement sur place
N : munir, démunir
P : verbes psychologiques
R : réalisation, mise en état
S : saisir, serrer, posséder
T : transformation, changement
U : union, réunion
X : verbes auxiliaires

Classes sémantico-syntaxiques Pour toutes les classes génériques, excepté C, D, P et X, les classes sémantico-syntaxiques encodent le type de sujet accompagnant le verbe ainsi que la distinction entre sens littéral et sens figuré.

- E, F, H, L, M, N, R, S, T, U
  1 : sujet humain ou animal, sens littéral
  2 : sujet humain, sens figuré
  3 : sujet inanimé, sens littéral
  4 : sujet inanimé, sens figuré

Pour les quatre autres classes génériques, la signification des codes sémantico-syntaxiques est différente.

- D (don, privation)
  1 : sujet humain
  2 : sujet non-humain, sens littéral
  3 : sujet non-humain, sens figuré

- C (communication)
  1 : sujet humain ou animal (crier, parler)
  2 : sujet humain (dire quelque chose)
  3 : sujet humain (montrer)
  4 : sens figuré

- P (verbes psychologiques)
  1 : sujet humain
  2 : objet humain
  3 : objet inanimé

- X (auxiliaires)
  1 : auxiliaires temporels ou aspectuels
  2 : impersonnels
  3 : synonymes de *être* + temps, lieu
  4 : *finir* et *commencer*

Sous-classes syntaxiques Elles sont au nombre de 248. Nous ne pouvons pas les détailler ici. À titre d'exemple, nous présentons les sous-classes correspondant à la

classe sémantico-syntaxique D1. Pour plus de détails sur cette sous-classification, nous renvoyons le lecteur à la documentation de Dubois & Dubois-Charlier (1997).

La classe D1 se divise en quatre sous-classes qui reflètent diverses constructions syntaxiques.

D1a : donner quelqu'un à quelque chose, se donner à quelque chose

(12) *la famille destine Paul à succéder à son père, on s'adonne au plaisir*

D1b : donner quelqu'un à quelqu'un, se donner à quelqu'un

(13) *les voisins ont dénoncé Paul à la police, on s'adjoint un collaborateur*

D1c : donner aide à quelqu'un (aider quelqu'un, aider à quelque chose)

(14) *on pistonne Paul auprès du directeur, on contribue à la réussite de ce projet*

D1d gratifier quelqu'un de quelque chose

(15) *on honore son pays avec cette victoire, le prêtre bénit l'assistance*

L'annotation selon les classes sémantiques de Dubois & Dubois-Charlier a été réalisée par les trois mêmes annotateurs que pour le caractère animé, en utilisant la version en ligne du dictionnaire [15] comme guide d'annotation. Nous avons à nouveau procédé en deux phases successives : une première phase d'annotation indépendante, puis une seconde de mise en commun et de prise de décision sur la base du consensus en cas de désaccord. Concrètement, nous avons ajouté une couche d'annotation au corpus de 1293 phrases annotées en syntaxe et pour le caractère animé. La principale difficulté posée par l'annotation des verbes réside dans la nature de la ressource utilisée. Le LVF n'a pas été construit dans l'optique d'une tâche d'annotation. Par conséquent, certains emplois rencontrés en corpus ne trouvent pas d'entrée dans le dictionnaire. À titre d'exemple, dans notre corpus, le verbe *mettre* est employé avec des SP prédicatifs, comme dans *mettre en valeur*. Cependant, le LVF n'a pas d'entrée pour ce type d'emploi. De plus, en tant que ressource construite à la main, le LVF propose parfois plusieurs entrées pour des sens proches d'un même verbe. Il est alors compliqué d'identifier ces nuances en contexte.

Les mesures d'accord pour le niveau de classification le plus large, classes génériques, sont satisfaisantes : Multi-π = 0.83 et Multi-κ = 0.85 (k=3, N=1434). Cela indique que les données relatives aux classes sémantiques générales sont fiables. La matrice de confusion en 6.10, construite sur le même principe que la table 6.8, présente les cas de désaccords un peu plus en détail.

En ce qui concerne les sous-classes syntaxiques, c'est-à-dire le niveau de classement le plus fin, l'accord inter-annotateur n'est pas satisfaisant : Multi-π = 0.72 et Multi-κ

15. http://rali.iro.umontreal.ca/Dubois/

	C	D	E	F	H	L	M	N	P	R	S	T	U	?
C	437	70	4	4	0	6	0	1	0	18	4	0	5	2
D		1360	39	4	10	26	8	15	35	40	38	4	9	20
E			485	1	6	5	8	0	3	0	12	2	11	17
F				27	0	0	0	4	0	0	0	0	0	1
H					95	6	0	0	1	6	3	0	0	0
L						445	9	2	17	39	29	0	1	6
M							136	0	0	3	0	2	2	12
N								55	0	0	8	0	4	0
P									58	2	1	0	7	2
R										97	0	2	0	2
S											61	0	5	0
T												191	2	4
U													237	4
?														5

TABLE 6.10.: Matrice de confusion pour les classes génériques des verbes de *TF* (*?* signifie 'Je ne sais pas')

= 0.73 (k=3, N=1434). Ces mesures sont largement inférieures au seuil généralement admis comme marquant un bon accord inter-annotateur (0.8). Cela signifie que les données annotées ne sont pas réellement fiables à ce niveau de précision dans la table *TF*. Ce mauvais accord s'explique par le nombre élevé d'étiquettes (248) et le fait que la ressource n'est pas réellement adaptée à la tâche d'annotation. Notons que, dans ce chapitre, notre analyse s'appuiera uniquement sur les classes génériques. Comme pour le caractère animé, les données concernant la sémantique du verbe utilisées dans l'analyse seront celles du corpus de référence.

6.2.2. Analyse

Nous procédons à l'analyse des données en deux étapes. Premièrement, nous présenterons les variables de la table *TF* captant les contraintes préférentielles que nous étudions, en détaillant la façon dont nous les avons obtenues ainsi que les premières observations que nous pouvons en faire. Deuxièmement, nous exposerons la modélisation de ces données.

6.2.2.1. Pronominalité

La table *TF* présente 8 occurrences de SP contenant un pronom personnel défini. Les deux ordres sont attestés comme en témoignent les phrases (16) et (17). On observe une tendance vers l'ordre SP SX_{OBJ}, avec 6 phrases, soit 75% des données, présentant cet ordre.

(16) *[…] a fait* ***de moi*** *son souffre-douleur* (ESTER)

(17) *[…] porte cette oeuvre* ***en lui*** (ESTER)

Si l'on ajoute aux pronoms personnels les pronoms démonstratifs *cela*, *ça*, *celui-ci* et toutes leurs variantes, nous obtenons un total de 14 SP et 9 SN pronominaux. Les proportions montrent une préférence pour l'ordre '*pronominal - non-pronominal*' : 100% pour les SN et 78.6% pour les SP. La tendance observée en anglais et en allemand semble être vérifiée pour le français, mais l'insuffisance du nombre de données ne permet pas de tirer de véritables conclusions. L'investigation de l'effet de cette contrainte devrait passer par la constitution d'un corpus spécifique, compilant un nombre beaucoup plus important de réalisations pronominales de compléments postverbaux. Étant donné le nombre réduit de données dont nous disposons, nous laissons de côté cette contrainte dans la suite de ce travail.

6.2.2.2. Contrainte de poids

En accord avec les résultats de l'étude préliminaire présentée dans la section précédente, nous mesurons le poids en nombre de mots. Pour capter l'effet de la contrainte de poids, nous utilisons la variable `longRelMots` qui se définit comme la différence des logarithmes de nombre de mots des deux constituants. L'utilisation des logarithmes permet de réduire l'intervalle sur lequel s'étendent les mesures de longueur et de réduire ainsi l'effet des valeurs extrêmes.

`longRelMots` : log(nombre de mots du SN) − log(nombre de mots du SP)

Les tendances observées dans la table préliminaire sont respectées. Le graphique de la figure 6.8 indique que plus le SN est court par rapport au SP, plus la proportion d'ordre SP SN a tendance à être basse et que, inversement, plus le SN est long, plus cette proportion est élevée. En suivant le principe *court avant long*, la variable `longRelMots` permet de connaître l'ordre attesté dans 82.5% des cas.

De plus, on observe que les 22 noms nus[16] de la table de données apparaissent directement après le verbe. Cela va dans le sens de la généralisation proposée par Abeillé & Godard (2004), selon laquelle les noms nus sont légers et doivent donc apparaître avant les autres constituants dans le domaine postverbal.

6.2.2.3. Séquence figée V SP

Dans certaines phrases, nous observons que le SP et le verbe forment une séquence figée non connexe. Il s'agit d'exemples tels que *mettre en valeur*, *mettre à disposition* ou *prendre en charge*. Malgré le caractère figé de ces séquences, on constate une possibilité d'alternance d'ordre. Le SP peut apparaître avant ou après le SN, comme en attestent les exemples (18) et (19).

16. Les noms nus de *TF* apparaissent dans les séquences suivantes : *faire peur*, *faire confiance*, *faire appel*, *donner asile*, *donner naissance*, *donner refuge*, *donner raison*, *donner droit*, *redonner espoir*, *laisser place*, *porter secours*, *prendre position*, *passer commande*.

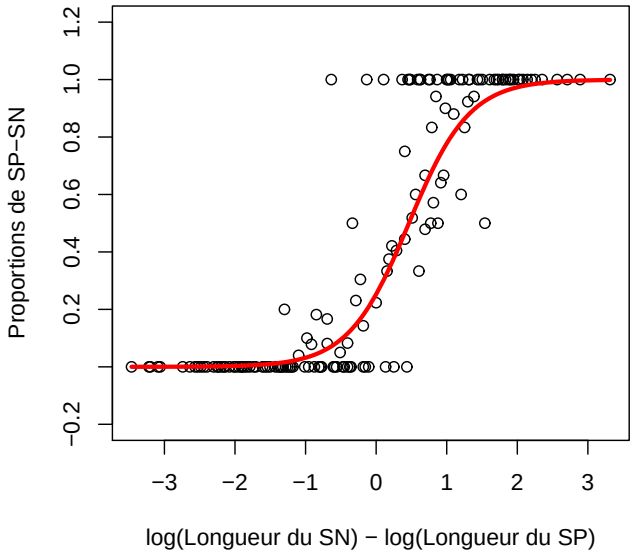

FIGURE 6.8.: La longueur relative des constituants (échelle logarithmique) en fonction de `ordre` avec la courbe logistique la mieux ajustée aux données.

(18) *il y avait déjà la confrérie la confrérie qui qui existe depuis des temps immémoriaux qui* ***prenait en charge*** *le mort parce que les gens payaient une cotisation à cette confrérie* (C-ORAL)

(19) *Le résultat est tout à fait remarquable et* ***met*** *l'ensemble du bâtiment* ***en valeur***. (ER)

Dans ces séquences, le SP participe au sens du prédicat, ce qui implique une "dépendance" sémantique entre les deux éléments. En effet, pour connaître le contenu exact du prédicat, il faut connaître le SP. En théorie, cette dépendance doit favoriser l'adjacence du verbe et du SP.

Nous avons d'abord repéré ces séquences figées sur la base d'un critère formel. Nous avons considéré comme potentiellement figés les SP composés d'une préposition et d'un nom sans déterminant. Nous avons ensuite trié manuellement les SP présélectionnés. Ainsi, nous avons écarté les séquences telles que *réduire de moitié* ou *ériger en succès*. Enfin, les séquences figées ont été repérées à l'aide de la variable `SPfige` :

`SPfige`

= 1 : le verbe et le SP forment une séquence figée,
= 0 : le verbe et le SP ne forment pas une séquence figée.

Les séquences figées ne représentent que 1.8%, soit 26 phrases de la table *TF*. La proportion d'ordre SP SN est de 69.2%. Étant donné que les séquences figées ne concernent que les SP de deux mots, on peut supposer que la proportion élevée

d'ordre SP SN est la conséquence de la contrainte de poids. Cependant, si l'on compare cette proportion avec celle concernant les SP de deux mots qui n'apparaissent pas dans une séquence figée, il apparaît que le facteur caractère figé s'ajoute à la contrainte de poids. En effet, parmi les 249 SP de deux mots non-figés, seuls 42.6% se présentent dans l'ordre SP SN. Le lien sémantique particulier qui unit le verbe et le SP semble donc favoriser l'ordre SP SN, au-delà de la contrainte de poids.

Comme pour le caractère pronominal, le nombre de données dont nous disposons est trop réduit pour avoir une réelle idée de l'impact de cette variable.

6.2.2.4. Caractère animé

Le caractère animé des référents est capté à l'aide de deux variables, `animSN` et `animSP`, dont les valeurs reposent sur la version de référence du corpus annoté manuellement [17].

`animSN`

= 1 : le référent du SN est animé,
= 0 : le référent du SN est inanimé ;

`animSP`

= 1 : le référent du SP est animé,
= 0 : le référent du SP est inanimé.

Premièrement, on observe que la majorité des données ont des référents inanimés : seuls 187 SN, soit 13%, et 578 SP, soit 40.3%, sont animés. La proportion de référents animés est donc plus élevée pour les SP. Deuxièmement, on constate des différences de proportions pour la variable `ordre` en fonction du caractère animé, comme cela est montré dans la table 6.11.

	SN SP		SP SN		Totaux	
TF	1010	70.4%	424	29.6%	1434	100%
`animSN` = 1	148	79.1%	39	20.9%	187	100%
`animSN` = 0	862	69.1%	225	30.9%	1247	100%
`animSP` = 1	379	65.6%	199	34.4%	578	100%
`animSP` = 0	631	73.7%	225	26.3%	856	100%

TABLE 6.11.: Le caractère animé en fonction de la variable `ordre` dans la table *TF*

D'après cette table, lorsque le référent du SN est animé, l'ordre SN SP est favorisé. Inversement, quand le SP est animé, c'est l'ordre SP SN qui présente une proportion plus élevée. Ces distributions sont statistiquement significatives : pour le SN, $\chi^2(1) = 7.3637$, $p < 0.01$; pour le SP, $\chi^2(1) = 10.601$, $p < 0.01$.

Étant donné que la contrainte générale relative au caractère animé est *animé avant inanimé*, il est intéressant d'observer les cas où le SP est animé alors que le SN ne

17. Rappelons que nous regroupons, sous l'étiquette *animé*, les catégories *humain*, *animal*, et *organisation*. Toutes les autres catégories se voient attribuer l'étiquette *inanimé* (cf. section 6.2.1.3).

l'est pas. Si le principe que nous venons d'énoncer est vérifié, la proportion d'ordre SP SN devrait augmenter pour les phrases présentant cette asymétrie. Parmi les 541 phrases pour lesquelles `animSN` = 0 et `animSP` = 1, l'ordre SP SN est représenté à 34%, ce qui est supérieur à la proportion générale. Il semble donc qu'il existe une tendance à placer les référents animés en premier. Cependant, dans la table *TPbis*, on observe que les variables `animSN` et `animSP`[18] n'ont pas d'effet significatif sur la distribution de `ordre`. Cette différence de comportement entre les deux tables laisse supposer que l'effet du caractère animé n'est pas central, ce qui sera confirmé dans la partie 6.2.2.7, lorsque les variables `animSN` et `animSP` seront intégrées à un modèle prenant en compte le poids des constituants et l'identité du lemme verbal.

6.2.2.5. Caractère défini

Le caractère défini du SN et du SP a été établi à partir des déterminants introduisant les SN. Ainsi, tout SN présentant un article défini, un déterminant démonstratif ou un déterminant possessif a été considéré comme défini. Dans tous les autres cas, il est indéfini. Un SP défini est un SP dans lequel l'objet de la préposition est un SN défini. Nous avons donc deux variables `SNdef` et `SPdef` qui encodent respectivement le caractère défini du SN et du SP.

`SNdef`

= 1 : le SN est défini,
= 0 : le SN est indéfini ;

`SPdef`

= 1 : le SP est défini,
= 0 : le SP est indéfini.

Les données relatives à ces deux variables sont présentées dans la table 6.12. On observe une légère influence du caractère défini du SN sur la variable `ordre`. En effet, parmi les 818 SN définis de cette table, 72.7% se présentent dans l'ordre SN SP, tandis que, pour les SN indéfinis, la proportion d'ordre SN SP est de 67.3%. Cette différence de distribution statistiquement significative ($\chi^2(1) = 4.6082$, $p < 0.05$) indique que les SN définis ont tendance à apparaître adjacents au verbe.

En anglais, la tendance peut se résumer par le principe *défini avant indéfini*. Afin de vérifier si ce principe est à l'oeuvre en français, nous avons isolé le sous-groupe de données dans lequel le SN est indéfini et le SP est défini. Si le principe est respecté, on devrait observer une proportion d'ordre SP SN plus importante que dans le reste du corpus. Parmi les 416 phrases pour lesquelles `SNdef`=0 et `SPdef`=1, la proportion d'ordre SP SN est de 33.2%, ce qui est supérieur à la moyenne générale du corpus (29.6%). Ces observations suggèrent que le principe *défini avant indéfini* s'applique en français et que le caractère défini du SP favorise légèrement l'ordre SP SN.

18. Les données de *TPbis* ont été annotées pour le caractère animé en suivant la même procédure que pour *TF*. Les mesures d'accord inter-annotateur sont Multi-π = 0.88 et Multi-κ = 0.92 (k=3, N=2284), pour l'opposition *animé* vs. *inanimé*.

	SN SP		SP SN		Totaux	
TF	1010	70.4%	424	29.6%	1434	100%
`SNdef` = 1	592	72.4%	226	27.6%	818	100%
`SNdef` = 0	414	67.2%	202	32.8%	616	100%
`SPdef` = 1	644	71.4%	258	28.6%	902	100%
`SPdef` = 0	362	68%	170	32%	532	100%

TABLE 6.12.: Le caractère défini en fonction de la variable `ordre` dans la table *TF*

L'examen plus précis des phrases dans lesquelles `SNdef` = 1 montre que parmi les définis, ce sont les SN possessifs qui ont un comportement particulier. Pour étudier cela, nous avons créé une nouvelle variable `SNpos` :

`SNpos`

= 1 : le SN est introduit par un déterminant possessif[19],
= 0 : le SN n'est pas introduit par un déterminant possessif.

Parmi les 195 SN introduits par un déterminant possessif, 83.6% apparaissent dans l'ordre SN SP. De plus, si l'on écarte les SN possessifs, on observe que l'ordre SN SP ne représente plus que 69.3% des données parmi les SN définis. Il semble donc que, dans *TF*, c'est le caractère possessif du SN qui a un impact sur le comportement de la variable `ordre`. Comme pour les autres contraintes, seule l'inclusion de ces variables dans un modèle général pourra nous donner une idée précise de leur rôle.

6.2.2.6. Lemme verbal et classe sémantique

Nous avons remarqué, dans l'étude préliminaire, que chaque verbe semble présenter une préférence pour un ordre ou pour l'autre. Les données de *TF* confirment cette observation.

L'annotation des verbes en classes sémantiques permet de distinguer les différents emplois verbaux. Pour capter ces différents emplois, nous avons créé une variable qui est la concaténation du lemme verbal et de sa classe générique annotée manuellement (cf. section 6.2.1.3).

`lemSem` : lemme verbal + classe générique

La liste présentée dans l'annexe F contient l'ensemble des valeurs de cette variable, accompagné des fréquences dans la table de données. Par exemple, pour le verbe *faire*, on distingue quatre emplois :

faire C (communication) : *Avec son excellent français, il se charge de* ***faire*** *la traduction à ses coéquipiers* (ER)

faire D (don/privation) : *La direction a veillé à* ***faire*** *un cadeau à toutes les femmes hospitalisées* (ER)

19. Les déterminants possessifs sont : *mon*, *ma*, *mes*, *ton*, *ta*, *tes*, *son*, *sa*, *ses*, *notre*, *nos*, *votre*, *vos*, *leur*, *leurs*.

faire R (réalisation) : *M. Netanyahu l'a accusé d'être prêt à* ***faire*** *des concessions inacceptables aux Palestiniens, reprenant ainsi les mêmes arguments qui lui avaient donné la victoire à l'arraché en 1996 contre les travaillistes.* (ER)

faire T (transformation) : *Finalement accepté moyennant des aménagements, il* ***fait*** *du groupe français le numéro un mondial en équipements de transmissions* (FTB)

Trois de ces emplois[20] correspondent à des préférences distinctes : alors que *faire R* et *faire C* favorisent à plus de 94% l'ordre SN SP, *faire T* ne s'observe qu'à 3.3% avec cet ordre. De la même façon, le verbe *mettre* connaît trois types d'emplois dans nos données :

mettre L (locatif) : *Jean-Pierre Chevènement a affirmé qu'il "ignorait" si le préfet Bonnet* ***avait mis*** *des documents dans un coffre d'une banque à l'étranger* (ER)

mettre D (don/privation) : *aujourd'hui, de nombreux tailleurs, bijoutiers, brodeurs ou stylistes,* ***mettent*** *leur savoir-faire à la portée de tous* (ESTER)

mettre R (réalisation) : *il y en a un, surtout un des 4, qui est persuadé qu'il y a quelque chose après la mort et qui* ***met*** *au point tout un système de signes à s'envoyer* (ESTER)

Pour deux de ces trois emplois, on observe deux tendances opposées : *mettre L* favorise l'ordre SN SP à 88.2%, tandis que *mettre R* a une préférence pour l'ordre inverse avec 63.2% de SP SN.

Les verbes et leurs emplois permettent d'observer des différences de préférence relative à l'ordre des compléments postverbaux. À titre d'exemple, nous présentons les fréquences d'ordre SN SP et SP SN pour trois valeurs de `lemSem` : *mettre L*, *vendre D* et *donner D*. La table 6.13 indique que *donner D* a une préférence proche de la préférence générale dans *TF*. Les verbes *mettre L* et *vendre D* ont un comportement distinct : *mettre L* présente une préférence plus forte pour l'ordre SN SP, tandis que *vendre D* favorise plutôt l'ordre inverse. Le graphique de la figure 6.9 montre que ces différences de comportement se retrouvent lorsque le poids relatif des constituants est pris en compte.

	`ordre` = 0		`ordre` = 1		Totaux	
mettre L	45	88.2%	6	11.8%	51	100%
donner D	67	73.6%	24	26.4%	91	100%
vendre D	11	36.7%	19	63.3%	30	100%

TABLE 6.13.: Comportement de la variable `ordre` en fonction de trois valeurs de la variable `lemSem` dans la table *TF*.

20. L'emploi *faire D* n'est représenté que par une seule occurrence dans la table de données.

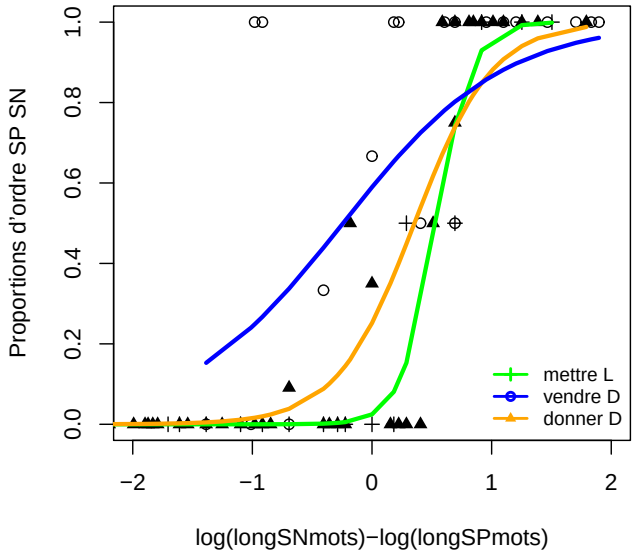

FIGURE 6.9.: La longueur relative des constituants (échelle logarithmique) en fonction de `ordre` pour *mettre L*, *vendre D* et *donner D*, avec les courbes logistiques résumant les données relatives à chaque verbe.

6.2.2.7. Modélisation

L'objectif de cette section est d'apprécier la significativité de chaque variable décrite précédemment en prenant en compte l'ensemble des variables envisagées. Pour cela, nous utiliserons la régression logistique à effets mixtes et les méthodes de comparaison de modèles. Les techniques statistiques utilisées sont présentées dans le chapitre 2.

Aspects "techniques" Nous modélisons la probabilité d'avoir l'ordre SP SN après le verbe, autrement dit la probabilité que `ordre` = 1. Le modèle de régression logistique à effets mixtes que nous utilisons se définit de la façon suivante :

$$P(\texttt{ordre} = 1|X, L_i) = \frac{e^{X\beta+L_i}}{1+e^{X\beta+L_i}} \tag{6.1}$$

où X renvoie aux variables prédictrices constituant les effets fixes et L_i aux effets aléatoires.

Nous avons observé, dans la section 6.2.1.2, une variation de proportion d'ordre SN SP selon le corpus, avec 67.8% dans FTB, 69.6% dans ER, 73.6% dans CORAL et 76.5% dans ESTER. Nous estimons donc que les données sont structurées autour de la variable `corpus`. Pour rendre compte de cela, nous considérons que cette variable constitue un effet aléatoire. Nous notons cet effet aléatoire C_i, où i représente le corpus considéré. Dans la modélisation, chaque corpus se voit attribuer un coefficient

propre qui rend compte de la distribution spécifique de `ordre` en son sein.

Nous construisons un Modèle Nul, qui contient l'effet aléatoire relatif à la variable `corpus`, mais qui ne contient aucune variable prédictrice. Le Modèle Nul prédit systématiquement l'échec (`ordre` = 0), c'est-à-dire l'ordre SN SP. Pour ce Modèle Nul, l'exactitude au seuil $P(\texttt{ordre} = 1|X) = 0.5$ est $E = 0.70$. L'aire sous la courbe ROC (AUC), qui est une autre mesure de qualité d'un modèle, est égale à 0.53. Ce modèle Nul servira de référence pour évaluer la qualité du modèle que nous allons présenter.

Ce modèle sera compacté sur la base du test de rapport de vraisemblance. Étant donné deux modèles imbriqués l'un dans l'autre, si le rapport de vraisemblance est statistiquement significatif, on considère alors que le modèle le plus complexe est justifié pour la modélisation de la variable `ordre`. En revanche, si le rapport de vraisemblance n'est pas significatif, alors on estime que le modèle le plus simple suffit à la modélisation.

Le modèle *TF* Le modèle est construit à partir de l'ensemble des variables prédictrices citées dans la section précédente : `longRelMots`, `SPfige`, `animSN`, `animSP`, `SNdef`, `SPdef` et `SNpos`. Étant donné qu'il existe des variations de distribution de la variable `ordre` selon les emplois verbaux, la variable `lemSem` est introduite comme un effet aléatoire [21], en plus de `corpus`. Le modèle a été compacté sur la base du test de rapport de vraisemblance. Tous les effets fixes ont été éliminés, excepté la longueur relative `longRelMots` ($\chi^2(6) = 5.9363$, p= 0.43). Le modèle est présenté dans la table 6.14. Le coefficient positif associé à la variable `longRelMots` indique que lorsque le SN est plus court que le SP, la variable vote pour l'ordre SN SP ; tandis que, lorsqu'il est plus long, la variable favorise l'ordre SP SN. De plus, l'intercept général du modèle est négatif, ce qui signale que le modèle vote pour l'ordre SN SP, si les deux constituants sont de la même longueur.

Les capacités de prédiction de ce modèle sont largement supérieures à celles du Modèle Nul : exactitude $\mu = 0.833$ ($\sigma = 0.034$) et $AUC = 0.901$ ($\sigma = 0.037$).

La variable `SPfige` n'est pas significative. Cela s'explique par le fait que très peu de données sont concernées par cette variable dans *TF*. Il semble que le caractère figé de la séquence V SP ne soit pas réellement significatif d'un point de vue quantitatif. En revanche, dans une approche plus qualitative, et comme le montrent les tendances observées dans *TF*, lorsqu'un lien sémantique particulier unit le verbe et le SP, l'adjacence du verbe et du SP est largement favorisée.

En ce qui concerne le caractère défini et possessif du SN, aucune des variables n'a d'apport significatif pour la modélisation de `ordre`. Cela va à l'encontre des premières

21. Nous introduisons le verbe en effet aléatoire sur l'intercept général du modèle. Étant donné l'allure des courbes de régression dans la figure 6.9, on pourrait penser que les verbes ont chacun un comportement différent par rapport à la variable de poids. Pour capturer un tel phénomène, il faudrait introduire un effet aléatoire par item lexical sur la pente associée à la variable de longueur. Cependant, dans le cadre de ce travail, le nombre d'observations par verbe étant relativement réduit, il semblerait artificiel de complexifier la structure des effets aléatoires du modèle. Pour introduire des effets aléatoires sur les pentes des variables, il serait nécessaire d'avoir plus d'observations par valeur de la variable `lemSem`.

```
Effets aléatoires :
  Groupes  Nom          Variance  Ecart-type
  lemSem   (Intercept)  1.24298   1.11489
  corpus   (Intercept)  0.24245   0.49239
Nombre d'obs. : 1434 ; groupes : lemSem, 253 ; corpus, 4

Effets fixes :
               Estimation  Erreur-type  valeur z  Pr(>|z|)
  (Intercept)  -1.4269     0.2879       -4.955    7.22e-07  ***
  longRelMots  2.6891      0.1565       17.183    < 2e-16   ***

Corrélation des effets fixes :
              (Intercept)
  longRelMots -0.128
```

Table 6.14.: Les paramètres du Modèle *TF*

observations que nous avions faites dans la section 6.2.2.5. Le graphique de la figure 6.10 semble indiquer que lorsque la longueur est prise en compte, l'effet de la variable `SNpos` est neutralisé.

Les intercepts aléatoires associés à chacun des corpus sont présentés dans la table 6.15. Les coefficients positifs indiquent un biais vers l'ordre SP SN, tandis que les négatifs signalent une préférence pour l'ordre inverse. On observe que les deux corpus journalistiques ont un intercept très proche. Les deux corpus oraux se distinguent l'un de l'autre : CORAL vote pour l'ordre SP SN alors que ESTER favorise l'ordre SN SP. Dans la mesure où les deux corpus oraux ont des préférences bien distinctes, il est donc difficile de tirer des conclusions en rapport avec le canal de transmission.

En revanche, on peut se demander si l'effet des corpus est en rapport avec l'échantillonnage. Rappelons que le corpus ER a été échantillonné en fonction de FTB, alors que ESTER et CORAL l'ont été de manière plus libre car il était plus difficile de trouver des occurrences pertinentes pour notre étude. Par exemple, CORAL est composé de 25 occurrences du verbe *mettre*, ce qui représente près de 23% des données de ce corpus dans *TF*. La proximité entre FTB et ER pourrait être due à une composition similaire en termes de lemmes verbaux. Si cette hypothèse se vérifie, cela accentuerait encore l'importance du lemme verbal dans le choix de l'ordre des compléments postverbaux.

CORAL	FTB	ER	ESTER
0.6025552	0.1362545	0.1627514	-0.6408192

Table 6.15.: Les intercepts aléatoires associés à la variable `corpus` dans la modèle *TF*.

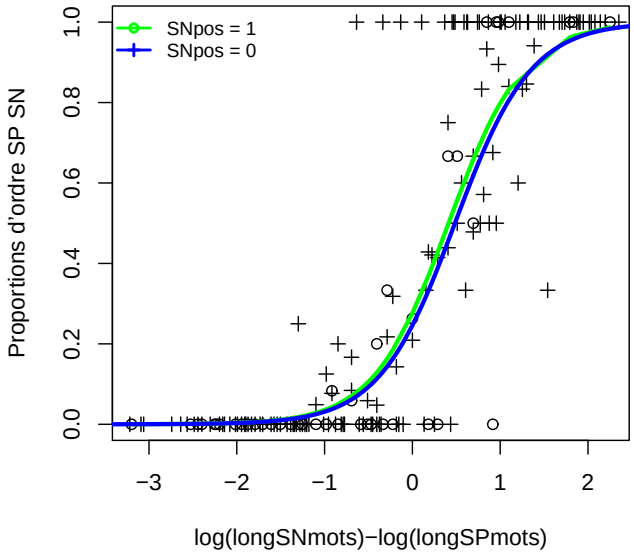

FIGURE 6.10.: La longueur relative des constituants (échelle logarithmique) en fonction de `ordre` pour les SN possessifs et les non-possessifs.

6.3. Verbe, caractère animé et statut du référent

Dans cette section, nous approfondissons l'étude de trois facteurs : le lemme verbal, le caractère animé et le statut informationnel des référents. Le premier facteur est important dans la mesure où notre étude a permis de le mettre à jour et de le formaliser. En ce qui concerne le deuxième facteur, les données modélisées semblent indiquer qu'il n'a pas d'impact sur l'ordre des constituants en français, ce qui va à l'encontre de ce qui a été observé dans les autres langues (cf. section 5.5.1). Enfin, le troisième facteur n'a pas été étudié dans nos données jusqu'ici. Nous donnons des éléments permettant d'évaluer son influence en français.

6.3.1. Biais verbaux et classes sémantiques

L'un des points majeurs émergeant de notre travail est le rôle joué par les verbes dans le choix de l'ordonnancement des compléments postverbaux. Grâce à la modélisation proposée, nous disposons d'une estimation des biais lexicaux. En effet, les intercepts aléatoires associés à chaque valeur de `lemSem` permettent d'avoir une bonne évaluation de l'influence de chaque verbe, dans la mesure où ces coefficients sont estimés dans un modèle prenant en compte l'effet du poids des constituants. Cependant, les valeurs des intercepts aléatoires sont dépendantes du modèle dans lequel elles sont estimées. Il est possible de comparer les intercepts aléatoires à l'intérieur d'un même modèle, mais nous ne pouvons pas comparer ces intercepts avec d'autres

valeurs calculées dans d'autres modèles.

Dans la figure 6.11, nous présentons les intercepts aléatoires pour les valeurs de `lemSem` ayant une fréquence supérieure à dix dans la table *TF*. L'intégralité des intercepts aléatoires est présentée dans l'annexe F. Comme pour les intercepts aléatoires associés aux corpus, un coefficient positif indique un élément favorisant l'ordre SP SN, et un intercept négatif un élément préférant l'ordre SN SP. Les barres présentes sur le graphique correspondent aux marges d'erreur sur l'estimation de chaque coefficient. Plus exactement, elles représentent l'intervalle de confiance à 95%. Cela signifie que la valeur estimée a 95% de chance de se trouver dans cet intervalle. Plus la barre d'erreur est réduite, plus l'estimation est fiable. Pour interpréter les valeurs de ces intercepts, il faut tenir compte de l'intercept général du modèle. Étant donné qu'il est de -1.426, le verbe doit avoir un intercept d'au moins +1.426 pour faire pencher la balance à lui seul vers l'ordre SP SN [22]. D'après le modèle, seuls sept verbes imposent l'ordre SP SN par défaut : *faire T*, *diminuer M*, *ajouter U*, *rendre D*, *écarter U*, *acheter D* et *obtenir S*. Pour tous les autres, c'est la combinaison avec la contrainte de poids ou le corpus d'origine qui peut amener à un vote pour l'ordre SP SN.

Nous avons montré que la classe générique associée au lemme verbal est un élément participant à la description de la distribution de la variable `ordre`. On peut maintenant s'interroger sur le rôle de la classe sémantique elle-même : existe-t-il un effet direct de la classe générique sur le choix de l'ordre des compléments ? Les données de la table 6.16 indiquent notamment que la classe T semble avoir une préférence pour l'ordre SP SN, la classe L pour l'ordre SN SP et que la classe D a un comportement proche de la moyenne générale. Ces données laissent supposer qu'il peut exister un effet lié à la classe générique telle que nous l'avons annotée.

Cependant, l'étude plus précise des lemmes composant ces classes indique une grande hétérogénéité de comportements à l'intérieur de chacune d'entre elles. Ainsi, en se reportant aux intercepts aléatoires donnés dans la figure 6.11, on observe que, pour chacune des classes, les différents verbes présentent des préférences opposées : *faire T* et *remplacer T*, *verser D* et *vendre D*, *mettre L* et *trouver L*. Il nous semble donc que les préférences doivent être envisagées au niveau du lemme verbal comme nous l'avons fait dans le Modèle *TF*. Néanmoins, la sémantique verbale peut faire partie des éléments façonnant la préférence de chaque lemme. En effet, nous concevons les biais verbaux comme le reflet d'un éventail de tendances s'exprimant à différents niveaux et s'imbriquant pour former des préférences spécifiques à chaque verbe. Nous apportons ici quelques éléments concernant la syntaxe et la sémantique du verbe qui peuvent être vus comme façonnant chaque biais lexical.

6.3.1.1. Niveau syntaxique

Wasow (1997) et Stallings *et al.* (1998) ont remarqué l'influence du cadre de sous-

22. Notons que cela est modulé en fonction du corpus : si la phrase est extraite de CORAL, la préférence doit être un peu moins forte, alors que si la phrase vient du corpus ESTER, le biais doit être encore plus fort.

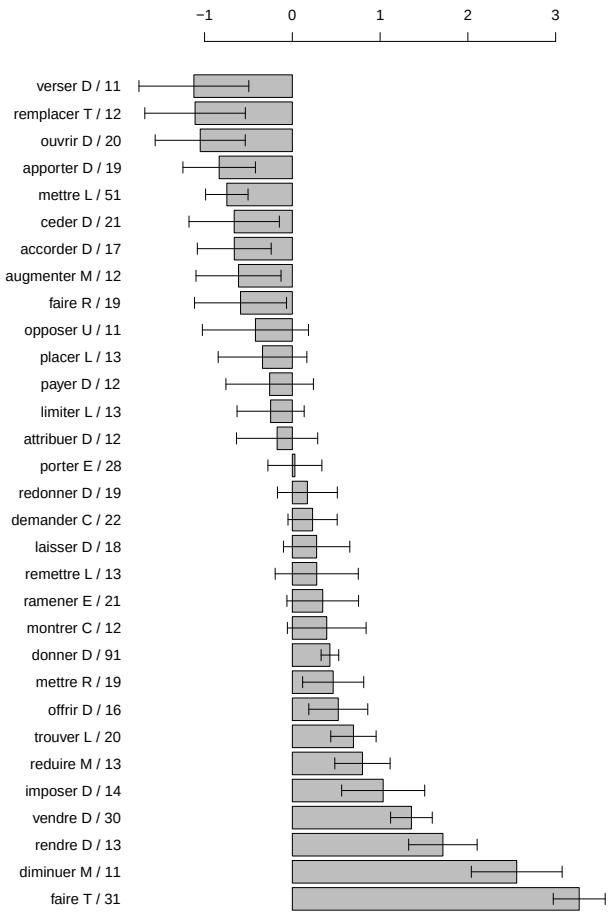

FIGURE 6.11.: Les intercepts aléatoires associés aux valeurs les plus fréquentes de `lemSem` dans le modèle *TF* ; le chiffre accompagnant chaque verbe correspond à la fréquence de ce dernier dans *TF*.

catégorisation des verbes sur l'ordre des mots en anglais (cf. section 5.5.2, chapitre 5). Nous émettons l'hypothèse qu'en français, le cadre de sous-catégorisation du verbe a également une influence sur la préférence de ce dernier.

Comme nous l'avons vu avec les données de l'étude préliminaire (table *TP*), lorsque l'objet est réalisé sous la forme d'une subordonnée ou d'une infinitive, l'ordre attesté est à plus de 99% V SP SX_{OBJ}. Ainsi, les verbes ayant la possibilité de réaliser leur objet sous cette forme sont presque systématiquement séparés de leur objet dans ces contextes. L'idée est alors que la fréquence de ces constructions a une influence sur la préférence générale du verbe. Plus exactement, nous nous inspirons de la *Verb disposition hypothesis*[23] formulée par Stallings *et al.* (1998) et nous l'adaptons au

23. « *individual verbs carry with them informations on the history of their participation in shifted structures and [...] this history influences the likelihood of their allowing HNPS* » (Stallings *et al.*,

	`ordre` = 0		`ordre` = 1		Totaux	
TF	1010	70.4%	424	29.6%	1434	100%
T(ransformation)	35	50.7%	34	49.3%	69	100%
L(ocatif)	137	82.5%	29	17.5%	166	100%
D(on)	365	69.7%	159	30.3%	524	100%
E(ntrée/sortie)	136	73.9%	48	26.1%	184	100%
F(rapper/toucher)	9	82%	2	18%	11	100%
H(états)	30	76.9%	9	23.1%	39	100%
M(ouvement)	26	52%	24	48%	50	100%
N(munir/démunir)	16	84.2%	3	15.8%	19	100%
P(sychologique)	27	81.8%	6	18.2%	33	100%
R(éalisation)	32	68.1%	15	31.9%	47	100%
S(aisir/serrer)	29	67.4%	14	32.6%	43	100%
U(nion/réunion)	72	80%	18	20%	90	100%

TABLE 6.16.: Comportement de la variable `ordre` en fonction des classes génériques dans la table *TF*.

problème qui nous intéresse : « *chaque verbe porte en lui des informations sur l'histoire de sa participation à des structures où l'objet direct est séparé du verbe par un* SP *complément et cette histoire influence la probabilité pour ce verbe d'apparaître dans des constructions où le* SN *objet direct est séparé du verbe par un* SP *complément (ordre* SP SN*)* ». Cela signifie que les verbes présentant plusieurs réalisations possibles pour leur objet direct (SN, subordonnée, infinitive) sont disposés à être séparés de ce dernier. D'après cette hypothèse, les verbes sous-catégorisant une subordonnée ou une infinitive objet ont un biais vers l'ordre SP SN.

Pour étayer cette idée, nous avons observé les préférences des verbes de la classe *Communication*[24]. Nous avons sélectionné les verbes ayant une fréquence strictement supérieure à 3, puis nous avons repéré ceux qui peuvent réaliser leur objet sous la forme d'une subordonnée ou d'une infinitive. La table 6.17 contient les verbes sélectionnés accompagnés de leur intercept aléatoire dans le Modèle *TF*.

L'ensemble des verbes appartenant à la classe *Communication* et sous-catégorisant une subordonnée objet ou une infinitive objet sont associés à un intercept positif dans le Modèle *TF*. Cela signifie qu'ils ont plutôt une préférence pour l'ordre SP SN. Ces données vont donc dans le sens de l'hypothèse que nous avons formulée. Notons toutefois que l'application de la *Verb disposition hypothesis* en français est à moduler dans la mesure où l'objet indirect est régulièrement réalisé sous la forme d'un clitique préverbal. Cela diminue le nombre de constructions dans lequel le verbe est séparé de son objet par l'objet indirect. Afin d'approfondir l'hypothèse sur l'influence du cadre de sous-catégorisation, il faudrait mener une étude plus détaillée sur ce type de verbes,

1998, p. 396).

24. L'ensemble des verbes appartenant à la classe *Communication* sont présentés avec leur fréquence et leur intercept aléatoire dans l'annexe F

lemSem	Intercept	Fréquence	Sub ou Inf
informer C	-0.381185801	4	non
faire C	-0.353060278	8	non
transmettre C	-0.254613864	4	non
donner C	-0.157798743	5	non
fixer C	-0.153532292	8	non
convaincre C	-0.059188444	5	non
prévenir C	0.014085105	5	non
expliquer C	0.147629476	5	oui
demander C	0.230851796	22	oui
présenter C	0.354769330	7	non
montrer C	0.392100207	12	oui
appeler C	0.464344657	5	non
dicter C	0.521238019	5	oui
confirmer C	0.745748810	4	oui
exiger C	0.766499931	4	oui
préconiser C	0.929562409	4	oui
apprendre C	1.150110185	4	oui
proposer C	1.184045730	10	oui

TABLE 6.17.: Intercepts aléatoires associés aux verbes de la classe C ayant une fréquence supérieure à 3 dans *TF*

en collectant un nombre de données beaucoup plus important par lemme. Il serait également pertinent de tester ces préférences à l'aide de méthodes expérimentales.

6.3.1.2. Niveau sémantique

Nous avons montré dans la section précédente que la classe sémantique, telle que nous l'avons annotée, n'a pas un comportement homogène et n'est donc pas pertinente pour l'étude. Il est possible qu'il existe des généralisations sémantiques, mais que les classes que nous avons utilisées n'en rendent pas compte. C'est ce que nous voudrions mettre en lumière en étudiant plus en détail les verbes appartenant à la classe *Transformation*. Dans la table 6.18, nous présentons ces verbes avec leur fréquence entre parenthèses et leur intercept aléatoire dans le Modèle *TF*.

compléter T (4)	convertir T (3)	échanger T (5)	ériger T (1)
-0.0828	-0.2361	+0.1439	-0.2082
faire T (31)	remplacer T (12)	transformer T (8)	troquer T (5)
+3.1616	-1.0393	-0.2064	+0.3407

TABLE 6.18.: Intercepts aléatoires associés aux verbes de la classe T avec la fréquence dans *TF* entre parenthèses.

Cinq de ces verbes peuvent être analysés comme exprimant une relation entre un patient (PAT) et un état résultant (RES) du patient : *compléter*, *convertir*, *ériger*, *transformer* et *faire*.

(20) Patient réalisé comme le SN

a. *[…]* ***compléter*** *[les assurances immeubles]$_{PAT}$ [par une renonciation à recours]$_{RES}$* (ER)
b. *[…]* ***convertir*** *[le pénalty]$_{PAT}$ [en but]$_{RES}$* (ER)
c. *[…]* ***ériger*** *[ses faiblesses]$_{PAT}$ [en succès]$_{RES}$* (FTB)
d. *[…] a* ***transformé*** *[les paisibles ruelles de la bourgade]$_{PAT}$ [en marché provençal]$_{RES}$* (ER)

(21) Patient réalisé comme le SP

a. *[…] ont* ***fait*** *[de cette journée]$_{PAT}$ [une grande réussite]$_{RES}$* (ER)

D'après les intercepts du Modèle *TF*, les quatre premiers verbes favorisent l'ordre SN SP tandis que le verbe *faire T* présente une nette préférence pour l'ordre SP SN. Ces préférences sont en correspondance avec l'ordre 'PAT $\prec$ RES. Nous émettons l'hypothèse que la relation établie entre les deux arguments par le verbe, a un impact sur l'ordre des compléments. Cela rejoint les observations de Schmitt (1987a,b) qui est, à notre connaissance, le seul auteur à avoir établi un lien entre l'ordre des arguments et le type de relation entretenue par ces arguments. Cependant, nous nous écartons de l'analyse de Schmitt dans la mesure où nous estimons que le lien sémantique entre les arguments et le verbe n'impose pas un ordre, mais se limite à le favoriser.

L'hypothèse que nous formulons concerne un groupe restreint de verbes pour lesquels le patient est affecté par la relation verbale et pour lesquels l'état résultant du patient est exprimé par le deuxième argument du verbe. Ce type de verbes peut être défini par rapport à l'échelle d'*affectedness* de Tsunoda (1985, p. 388). Cette échelle permet de classer les verbes selon les degrés auxquels le prédicat verbal affecte son patient. Les cinq verbes auxquels nous nous intéressons sont en haut de l'échelle de Tsunoda. Il s'agit de verbes exprimant une action directe sur le patient et pour lequel on obtient un résultat.

Soit β et γ les deux arguments du verbe et R(β, γ) la relation qu'instaure le prédicat entre ses deux arguments, si β est affecté par R(β, γ) et que γ correspond à l'état résultant de β, alors les arguments ont tendance à apparaître linéairement dans l'ordre β γ.

Pour les verbes *compléter*, *convertir*, *ériger* et *transformer*, le patient est réalisé comme un objet direct et l'état résultant comme un oblique. L'ordre attesté au niveau syntaxique est donc SN$_{OBJ}$ SP$_{OBL}$. En revanche, dans le cas de *faire T*, le patient est exprimé sous la forme du SP et l'état résultant sous la forme du SN objet. L'ordre observé au niveau des réalisations syntaxiques est donc SP$_{OBL}$ SN$_{OBJ}$. En conclusion, ces verbes introduisent une relation entre les arguments β et γ qui influence l'ordre

des deux arguments, indépendamment de leur réalisation syntaxique [25].

Nous avons esquissé une analyse pour un petit groupe de verbes, permettant d'expliquer les biais verbaux. Ce cas particulier ouvre une piste de recherche plus large sur l'étude de l'ordre des compléments en fonction de la relation sémantique entre le verbe et ses deux arguments. Il faudrait notamment considérer l'impact de la hiérarchie des rôles sémantiques sur l'ordonnancement des compléments (cf. section 5.5.2, chapitre 5).

6.3.2. Le caractère animé

Le modèle *TF* montre que le caractère animé des référents ne participe pas significativement à la modélisation de la variable `ordre`.

Les variables relatives au caractère animé ont été annotées manuellement et ont fait l'objet de désaccords, comme nous l'avons mentionné dans les sections traitant de l'annotation (cf. section 6.2.1.3). Afin de nous assurer que la non-significativité du caractère animé n'est pas dû à un problème d'annotation, nous avons étudié une sous-partie de la table, que nous appelons *TDanim*, pour laquelle le statut animé ou inanimé des référents n'était pas problématique. Nous avons sélectionné 1000 phrases pour lesquelles les trois annotateurs étaient en accord sur le caractère animé ou inanimé des référents du SN et du SP. Les proportions de SN animés, de SP animés et d'ordre SN SP dans *TDanim* sont équivalentes à celles de *TF*. On observe également que la distribution de la variable `ordre` en fonction de `animSN` et `animSP` est très proche de celle observée dans *TF* [26]. Si l'on construit un modèle pour rendre compte du comportement de la variable `ordre` en fonction des effets fixes `logRelMots`, `animSN`, `animSP` et des effets aléatoires `corpus` et `lemSem`, on observe, par comparaison de modèles, que les variables `animSN` et `animSP` ne sont pas significatives ($\chi^2(2) = 0.7614$, p= 0.68). Ainsi, les données présentant une annotation totalement fiable aboutissent au même résultat que les données de l'ensemble de la table *TF*.

On peut se demander pourquoi les variables `animSN` et `animSP` qui, d'après les premières observations, avaient une influence sur l'ordre des compléments, s'avèrent ne pas être pertinentes dans le cadre du modèle de prédiction. Les deux dimensions aidant à la description du comportement de la variable `ordre`, sont le poids relatif des constituants et l'identité du lemme verbal en contexte. Les relations entretenues entre `longRelMots`, `lemSem` et les deux variables relatives au caractère animé devraient donc expliquer la différence entre les premières observations et la modélisation.

Premièrement, intéressons-nous au comportement des variables `animSN` et `animSP` pour les phrases dans lesquelles les deux constituants ont la même longueur. Les proportions présentées dans la table 6.19 indiquent une légère préférence pour l'ordre

25. Nous tenons à remercier Jean-Marie Marandin pour son aide dans l'analyse de ces données.

26. Les chiffres exacts sont : 70.7% d'ordre SN SP, 13.3% de SN animés et 43.4% de SP animés. Les proportions d'ordre SN SP en fonction des variables `animSN` et `animSP` sont les suivantes : 78.2% pour les SN animés, 69.5% pour les SN inanimés, 65.1% pour les SP animés et 75.0% pour les SP inanimés.

SN SP quand les référents des SN et des SP sont animés. Cependant, ces différences ne sont pas significatives : pour `animSN`, $\chi^2(1) = 8 \times 10^{-4}$, $p = 0.98$; pour `animSP`, $\chi^2(1) = 0.0358$, p= 0.85. De plus, le graphique de la figure 6.12 indique qu'une fois pris en compte le poids des constituants, la variable `ordre` ne montre pas de nettes différences de comportement en fonction des valeurs des variables `animSN` et `animSP`. Il apparaît donc que les effets du caractère animé dans *TF* sont largement liés à la longueur des constituants.

	SN SP		SP SN		Totaux	
`animSN` = 1	24	80%	6	20%	30	100%
`animSN` = 0	144	77.8%	41	22.2%	185	100%
`animSP` = 1	62	79.5%	16	20.5%	78	100%
`animSP` = 0	106	77.4%	31	22.6%	137	100%

TABLE 6.19.: Le caractère animé en fonction de la variable `ordre` pour les phrases où `longRelMots` = 0.

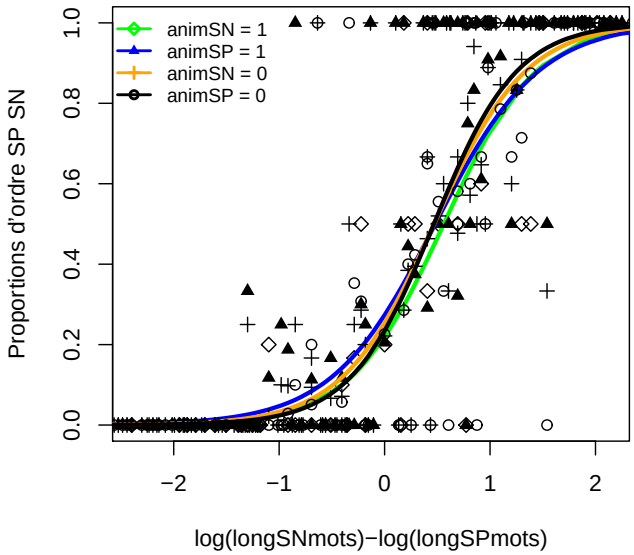

FIGURE 6.12.: La longueur relative des constituants (échelle logarithmique) en fonction de `ordre` pour les variables `animSN` et `animSP` avec les courbes logistiques résumant les données relatives.

Deuxièmement, le caractère animé est en lien avec l'identité du lemme verbal et avec sa sémantique. En effet, pour des verbes de communication, tels que *demander*, ou de don, tels que *vendre*, le SP a un rôle sémantique de bénéficiaire ou destinataire

qui est généralement rempli par un référent animé. Pour les deux verbes que nous avons cités, la proportion de SP animés est supérieure à 90%. Par ailleurs, ces deux verbes ont une préférence pour l'ordre SP SN, à 63.3% pour *vendre D* et à 40.9% pour *demander C*. De façon plus générale, la distribution de `animSP` est corrélée à la nature du lemme verbal en contexte. Pour les dix valeurs de `lemSem` les plus fréquentes dans *TF*[27], la corrélation est statistiquement significative : $\chi^2(9) = 167.5145$, $p < 0.0001$. La sémantique du verbe est donc largement liée au caractère animé du SP dans *TF*. Ce constat nous amène à émettre l'hypothèse que le caractère animé peut être un aspect sémantique façonnant les biais verbaux (cf. section précédente).

Une fois mises à jour les corrélations entre poids, identité du lemme verbal et caractère animé, nous pouvons nous interroger sur le véritable effet du caractère animé en français. En effet, dans les travaux cités dans la section 5.5.1, ce caractère apparaît comme un facteur déterminant l'ordre des arguments verbaux dans des langues telles que l'anglais, le grec ou l'allemand. Ces observations amènent certains auteurs à affirmer que le caractère animé est un facteur universel. Nos résultats tendent à montrer que cela ne se vérifie pas pour l'ordre des compléments postverbaux en français.

6.3.2.1. Élicitation de jugements d'acceptabilité pour le caractère animé

Étant donné que le matériel extrait des corpus contient des corrélations difficiles à neutraliser, on pourrait supposer que, dans nos données, l'effet du caractère animé est masqué par celui d'autres variables. Pour tester cette idée, nous avons mis en place, en collaboration avec Anne Abeillé et Benoît Crabbé, un questionnaire psycholinguistique avec l'objectif de recueillir les jugements de locuteurs natifs sur l'ordre des compléments en fonction du caractère animé[28].

L'hypothèse de départ est que les locuteurs ont tendance à préférer l'ordre *animé avant inanimé*. Le questionnaire a été conçu pour observer cette préférence au moyen de jugements d'acceptabilité sur des phrases dans lesquelles les autres variables sont neutralisées.

Le questionnaire est composé de 16 phrases présentant un verbe suivi de deux compléments sous-catégorisés. Dans ces phrases, le SN et le SP ont la même longueur en nombre de mots et sont tous deux définis. De plus, les lemmes verbaux ont été choisis en fonction de leur préférence observée en corpus, afin de varier au maximum les biais qu'ils imposent. En ce qui concerne le caractère animé, le référent du SN est toujours inanimé et seul le SP varie pour ce critère. Si l'hypothèse de départ est vérifiée, alors on devrait observer une préférence des locuteurs pour l'ordre SP_{anim} SN_{inanim}.

Chacune des 16 phrases est présentée avec les deux ordres possibles. Les sujets

27. Nous n'avons retenu que les dix verbes en contexte les plus fréquents, afin d'éviter la dispersion des données. Les dix lemmes sélectionnés sont : *ouvrir D*, *trouver L*, *céder D*, *ramener E*, *demander C*, *porter E*, *vendre D*, *faire T*, *mettre L*, *donner D*.

28. Ce travail est la prolongation d'une première étude présentée à la conférence Architectures and Mechanisms for Language Processing (Thuilier *et al.*, 2011).

doivent exprimer un jugement d'acceptabilité sur une échelle de Likert (échelle à 5 points) pour chacune des alternatives, comme cela est montré dans l'exemple (22). La note 1 correspond à *pas acceptable* et la note 5 à *complètement acceptable*.

(22) *En Camargue, seuls les flamants roses peuvent*
donner à un paysage monotone une pointe de relief. **1** ☐ ☐ ☐ ☐ ☐ **5**
donner une pointe de relief à un paysage monotone. **1** ☐ ☐ ☐ ☐ ☐ **5**

En plus des 16 phrases concernant le caractère animé, le questionnaire contient 22 distracteurs. Pour chaque questionnaire, l'ordre dans lequel sont présentées les 38 phrases est déterminé aléatoirement. Un exemple de questionnaire est présenté dans l'annexe C. Trente-huit sujets, étudiants L2 de linguistique à l'Université Paris Diderot, ont complété le questionnaire.

Les premières observations graphiques (à gauche sur la figure 6.13) montrent que les jugements des locuteurs sont sensibles à l'ordre des constituants : l'ordre SN SP reçoit un jugement moyen de 4.06 points tandis que l'ordre SP SN ne reçoit qu'une moyenne de 3.68 points. Le graphique du centre donne les jugements moyens selon le statut du SP. La moyenne pour les SP animés semble plus élevée. Néanmoins, les barres d'erreur indiquent que les deux moyennes ne sont pas significativement différentes. Le dernier graphique, à droite, donne une représentation de l'interaction des variables `ordre` et `animSP`. Il permet d'observer la moyenne des jugements pour les couples de valeurs suivants : 'SNSP-SPanim', 'SNSP-SPinanim', 'SPSN-SPanim' et 'SPSN-SPinanim'. On retrouve la différence de jugement entre les ordres SN SP et SP SN, quelle que soit la nature du référent du SP. On observe également que le couple 'SPSN-SPanim' obtient en moyenne une note supérieure à celle attribuée à 'SPSN-SPinanim', ce qui va dans le sens de notre hypothèse de départ.

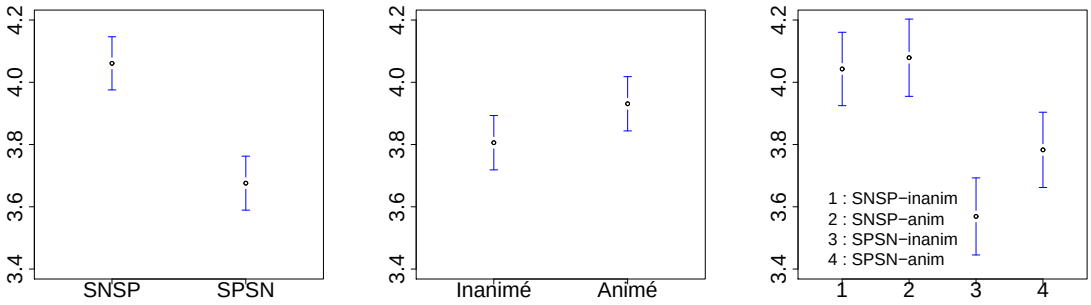

FIGURE 6.13.: À gauche, moyenne des jugements en fonction de l'ordre des compléments verbaux ; au centre, moyenne des jugements en fonction du caractère animé du SP ; à droite, moyenne des jugements en fonction de l'interaction entre ordre et caractère animé du SP.

Nous avons modélisé les résultats de l'expérience en utilisant un modèle linéaire à

effets mixtes (cf. section 2.2.1.3, chapitre 2) [29]. Le modèle a pour objectif de rendre compte des jugements des locuteurs, encodés par la variable `jug`. Afin de tenir compte des variations de jugements selon les sujets et selon les phrases de l'expérience, nous avons traité ces deux dimensions comme des effets aléatoires (S_i, P_j) [30]. Nous avons construit le modèle autour des effets fixes `ordre`, `animSP` et de l'interaction entre `ordre` et `animSP`. Après réduction de modèle sur la base du test du rapport de vraisemblance, la variable `animSP` et l'interaction sont écartées du modèle. Seule la variable `ordre` est conservée. Le modèle est présenté dans la table 6.20.

```
Effets aléatoires :
  Groupes      Nom          Variance  Ecart-type
  Sujet        (Intercept)  0.136474  0.36942
  LemmeVerbal  (Intercept)  0.016882  0.12993
  Résidus                   1.023341  1.01160
Nombre d'obs. : 1216 ; groupes : Sujet, 38 ; LemmeVerbal, 7

Effets fixes :
               Estimation  Erreur-type  valeur t  Pr(>|z|)
  (Intercept)  4.02447     0.08917      45.13
  ordre=1      -0.38487    0.05802      -6.63

Corrélation des effets fixes :
           (Intercept)
  ordre=1  -0.325
```

TABLE 6.20.: Modélisation des résultats du questionnaire concernant l'effet du caractère animé du SP sur les jugements des locuteurs natifs.

D'après ce modèle, l'ordre SP SN fait diminuer le jugement des locuteurs (coefficient négatif). Le modèle ne nous fournit aucun élément allant dans le sens de notre hypothèse de départ, à savoir que l'ordre SP_{anim} SN_{inanim} est préféré.

Afin de nous assurer que le caractère animé du SP n'influe pas sur les préférences des sujets, nous avons pris la sous partie des données pour laquelle l'ordre est SP SN. Dans ce sous-ensemble, seul le caractère animé du SP varie. Nous cherchons à voir si les jugements fluctuent en fonction de la variable `animSP` dans ce contexte particulier. Le modèle linéaire à effets mixtes contient les deux effets aléatoires relatifs aux sujets et aux phrases, ainsi que la variable prédictrice `animSP`. Après comparaison de modèles, il apparaît que la variable `animSP` n'est pas significative dans ce contexte.

29. Le choix de ce type de modélisation est partiellement problématique. En effet, nous cherchons à modéliser des scores compris entre 0 et 5 avec un modèle linéaire pour lequel la variable à prédire peut prendre des valeurs dans l'intervalle $]-\infty, +\infty[$. Nous estimons cependant que ce modèle est une approximation correcte pour traiter les résultats de notre expérience.

30. Nous traitons les phrases de l'expérience comme un effet aléatoire afin de prendre en compte le fait que chaque item présenté aux sujets introduit un biais vers un jugement plutôt positif ou plutôt négatif.

La modélisation des jugements portant sur les phrases à ordre SP SN montre donc qu'un SP animé adjacent au verbe ne recueille pas de meilleurs jugements qu'un SP inanimé.

Les résultats de ce questionnaire ne permettent pas d'affirmer que l'hypothèse de départ, selon laquelle les locuteurs ont tendance à préférer l'ordre *animé avant inanimé*, n'est pas vraie. Néanmoins, les données de corpus et les jugements d'acceptabilité vont dans le même sens et tendent à montrer que le caractère animé n'a pas d'effet sur l'ordre des compléments postverbaux en français. Comme nous l'avons signalé précédemment, le caractère animé des constituants est en lien avec la sémantique du verbe. Par conséquent, l'effet du caractère animé n'existe peut-être pas de façon directe, mais plutôt à travers la sémantique du verbe et le rôle sémantique de ses arguments. Cette idée rejoint l'hypothèse *indirecte* (cf. section 5.5.1 du chapitre précédent) défendue par McDonald *et al.* (1993) : l'influence du caractère animé ne s'exprime pas directement dans la linéarisation des compléments, mais elle joue un rôle au niveau de l'assignation des fonctions grammaticales.

Dans la section 5.5.4, nous avions mentionné que l'ordonnancement des compléments postverbaux était intéressant pour tester la validité de l'hypothèse *directe* en français. D'après cette hypothèse, le caractère animé a une influence sur l'ordre des dépendants verbaux indépendamment de l'assignation des fonctions grammaticales. (Branigan & Feleki, 1999; Branigan *et al.*, 2008; Kempen & Harbusch, 2004; Tanaka *et al.*, 2011). Les résultats obtenus sur corpus et dans le questionnaire ne fournissent pas d'argument en faveur de cette hypothèse dans la mesure où l'ordre et les préférences des locuteurs n'apparaissent pas directement affectés par le caractère animé des référents. Cependant, afin de confirmer la non-pertinence du caractère animé, il conviendrait de mener des expériences telles que celles proposées par Branigan & Feleki (1999) et Tanaka *et al.* (2011) (répétition de phrases préalablement entendues, cf. section 5.5.1 du chapitre 5).

6.3.3. L'opposition *donné* vs. *nouveau*

Le modèle que nous avons construit à partir des données de la table *TF* ne permet de prédire qu'environ 83% des données avec un seuil fixé à 0.5 et sa capacité de classification est autour de 0.90 (mesure *AUC*). Cela signifie que la modélisation de la variable `ordre` peut être améliorée et qu'il existe probablement d'autres contraintes préférentielles que nous n'avons pas prises en compte.

Nous nous sommes intéressée à l'effet du statut du référent du SN et du SP selon les catégories de Prince (1981). Cet auteur a proposé une classification des référents en fonction de l'accessibilité de ces derniers dans le discours. Cette typologie repose sur l'idée selon laquelle les entités de discours sont répertoriées dans un MODÈLE DE DISCOURS, qui est "partagé" par les interlocuteurs. Nous avons retenu les trois principales catégories de Prince : 'nouveau' (*new*), 'inférable' (*inferrable*), 'évoqué' (*evoked).* Ces catégories sont définies comme suit :

nouveau « *quand le locuteur introduit pour la première fois une entité dans le discours, nous pouvons dire qu'elle est* NOUVELLE » [31]. Il existe deux sous-types : d'une part, les *toutes nouvelles* entités qui sont créées par le locuteur dans le but de les introduire dans le modèle de discours ; d'autre part, les entités *inutilisées* qui sont en général connues par l'interlocuteur. Elles ne doivent pas être créées, mais simplement introduites dans le modèle de discours.

évoqué : « *si un* SN *est produit alors qu'il est déjà dans le modèle de discours […] il représente une entité* ÉVOQUÉE. » [32]. Les référents peuvent être évoqués textuellement ou situationnellement.

inférable : « *une entité de discours est Inférable si le locuteur peut l'inférer via un raisonnement logique − ou plus communément plausible −, à partir des entités de discours déjà Évoquées ou à partir d'autres Inférables* » [33].

D'après ce qui a été observé dans les autres langues, le principe général est que ce qui est *évoqué* (ou *donné*) a tendance à apparaître avant ce qui est *nouveau*.

Nous avons étudié l'influence du statut des référents du SN et du SP dans le discours, sur l'ordre des compléments postverbaux en français. Pour cela, nous avons d'abord annoté une sous-partie de la table *TF* en fonction de ces catégories, puis nous avons mis en place un questionnaire afin d'évaluer si les jugements des locuteurs étaient influencés par le statut des référents dans le discours.

6.3.3.1. Statut des référents en discours dans *TF*

Nous avons établi un échantillon de phrases pour lesquelles la différence de longueur entre les deux compléments ne dépasse pas deux mots. Nous avons ainsi cherché à limiter l'effet du poids pour pouvoir observer l'effet du statut des référents dans le discours. Les 204 phrases sélectionnées ont été annotées manuellement par l'auteur pour le statut du SN et du SP selon les trois catégories précédemment définies [34]. L'étiquette *inférable* a été utilisée 13 fois pour des SN et 27 fois pour des SP. Étant donné que cette catégorie est la plus difficile à définir et qu'elle ne représente que peu de données, nous avons décidé d'écarter les phrases contenant des *inférables*, pour ne garder que les 166 phrases incluant des référents *donnés* et *nouveaux*. Dans ce sous-ensemble, la proportion d'ordre SN SP est de 79.5%. Nous créons deux variables, `statutSN` et `statutSP`, qui encodent le statut des référents des compléments verbaux.

31. « *When a speaker first introduces an entity into the discourse […] we may say that it is* NEW », (Prince, 1981, p. 235).

32. « *if some NP is uttered whose entity is already in the discourse model […] it represents an* EVOKED *entity* », (Prince, 1981, p. 236).

33. « *a discourse entity is Inferrable if the speaker assumes the hearer can infer it, via logical − or, more commonly, plausible − reasonning, from discourse entities already Evoked or from other Inferrables.* », (Prince, 1981, p. 236).

34. L'annotation n'a pas été faite sur l'ensemble du corpus par plusieurs annotateurs. Les données dont nous disposons sont donc moins fiables que pour le caractère animé et pour les classes sémantiques verbales. Néanmoins, nous pensons qu'elles peuvent permettre de mettre à jour une corrélation entre statut des référents et ordre des compléments, si toutefois elle existe.

`statutSN`
= 1 : le référent du SN est donné,
= 0 : le référent du SN est nouveau ;

`statutSP`
= 1 : le référent du SN complément de la préposition introduisant le SP est donné,
= 0 : le référent du SN complément de la préposition introduisant le SP est nouveau.

Les données relatives à ces deux variables sont présentées dans la table 6.21. D'après les proportions observées, un SN *donné* a tendance à favoriser l'ordre SN SP par rapport à un SN *nouveau*. Pour le SP, les données suivent aussi ce qui est attendu, à savoir que le SP *donné* favorise l'ordre SP SN, mais la tendance est plus faible.

	SN SP		SP SN		Totaux	
SN donné	37	88.1%	5	11.9%	42	100%
SN nouveau	95	76.6%	29	23.4%	124	100%
SP donné	46	78%	13	22%	59	100%
SP nouveau	86	80.4%	21	19.6%	107	100%

TABLE 6.21.: La variable `ordre` en fonction de `statutSN` et `statutSP` dans une sous partie de *TF*.

Ces premières observations sont encourageantes mais ne sont pas significatives : pour les SN, $\chi^2(1) = 1.8836$, p= 0.17 et pour les SP, $\chi^2(1) = 0.0279$, p= 0.87. De plus, si l'on construit un modèle de régression logistique avec le lemme verbal et le corpus comme effets aléatoires, alors les variables `statutSN` et `statutSP` sont éliminées du modèle lorsque ce dernier est compacté. Les résultats obtenus à partir des données annotées ne sont donc pas concluants en ce qui concerne l'effet du statut des compléments en discours.

6.3.3.2. Élicitation de jugements d'acceptabilité pour l'opposition *donné/nouveau*

De la même façon que pour le caractère animé, on peut supposer que les corrélations présentes dans les données de corpus ne permettent pas d'identifier l'influence des contraintes `statutSN` et `statutSP`. Dans le but d'observer l'effet du statut informationnel des référents dans des données où les corrélations sont contrôlées, nous avons mis en place un questionnaire visant à tester l'effet du statut *donné* ou *nouveau* du SP sur le jugement des locuteurs. Le questionnaire est bâti sur le même principe que celui élaboré pour l'étude du caractère animé (cf. section 6.3.2.1).

Il contient 16 phrases dans lesquelles la longueur des constituants est neutralisée et où le SN est toujours nouveau et indéfini. Huit lemmes verbaux ont été sélectionnés en fonction de leurs préférences calculées en corpus. Seul le statut du SP varie : il

est donné et défini dans la moitié des phrases ; il est nouveau et indéfini dans l'autre moitié. Pour chaque verbe, il existe donc un contexte où le SP est donné et un autre où il est nouveau. Les phrases sont présentées avec deux continuations possibles. Les sujets doivent juger l'acceptabilité de chaque phrase en utilisant une échelle allant de 1 à 10 (1 = *pas acceptable*, 10 = *complètement acceptable*) [35]. Deux exemples sont présentés en (23) et (24) : dans le premier, le SP est nouveau ; dans le second, il est donné.

(23) *En Camargue, seuls les flamants roses peuvent*
donner à un paysage monotone une pointe de relief. **1** □ □ □ □ □ □ □ □ □ □ **10**
donner une pointe de relief à un paysage monotone. **1** □ □ □ □ □ □ □ □ □ □ **10**

(24) *De nombreuses questions se posent à propos de la situation économique du pays. Il faut que les candidats maintenant*
donnent à ces questions des réponses appropriées. **1** □ □ □ □ □ □ □ □ □ □ **10**
donnent des réponses appropriées à ces questions. **1** □ □ □ □ □ □ □ □ □ □ **10**

L'hypothèse de départ étant qu'il existe une préférence pour l'ordre *donné avant nouveau*, on s'attend à observer une préférence des locuteurs pour l'ordre SP_{donne} $\text{SN}_{nouveau}$. Aux 16 phrases concernant le statut du SP s'ajoutent 17 distracteurs. Pour chaque formulaire, l'ordre d'apparition des phrases a été randomisé. Le questionnaire a été rempli par 28 sujets, locuteurs du français, âgés de 17 à 27 ans et étudiants à l'Université Paris Sorbonne. L'annexe D contient un exemplaire du questionnaire.

Les moyennes des jugements recueillies pour les 16 phrases sont présentées dans les graphiques de la figure 6.14. On observe que la moyenne des jugements est plus élevée pour l'ordre SN SP que pour l'ordre SP SN (graphique de gauche). En revanche, le statut du SP ne semble pas influer sur la moyenne des jugements (graphique au centre). Enfin l'interaction de l'ordre et du statut du SP ne fait pas apparaître d'autre effet que celui de l'ordre.

Ces premiers résultats sont confirmés par la modèle linéaire à effets mixtes construit pour rendre compte des jugements des locuteurs. Pour tenir compte des différences de jugements selon les sujets et selon les phrases de l'expérience, nous mettons les sujets et les phrases en effets aléatoires. Les deux variables prédictrices du modèle sont l'ordre et le statut du SP, ainsi que l'interaction de ces deux variables. Après avoir compacté le modèle sur la base du test du rapport de vraisemblance, la variable `statutSP` et l'interaction sont éliminées, car elle ne participent pas significativement à la modélisation des jugements ($\chi^2(2) = 0.3965$, p= 0.82). Les résultats du questionnaire ne permettent donc pas de tirer de conclusion en ce qui concerne l'effet du statut du SP sur les jugements. Deux hypothèses peuvent expliquer ce résultat :

A. il n'existe pas d'effet et c'est pour cela que l'on n'en observe pas ;

B. il existe un effet et notre questionnaire ne permet pas de l'observer.

Le questionnaire que nous avons mis en place avait pour but de vérifier l'existence

35. Pour cette deuxième expérience, nous avons opté pour une échelle allant de 1 à 10, car il semble que les français soient plus habitués à utiliser une échelle de 10 qu'une échelle de 5, notamment en raison de l'utilisation de ce type d'échelle pour la notation scolaire.

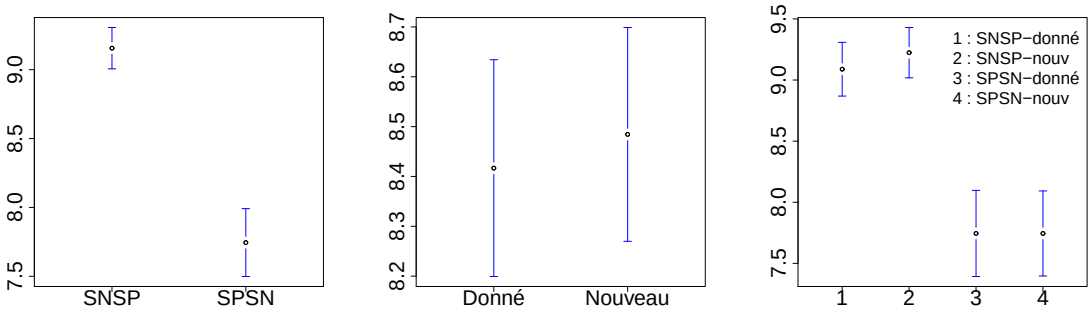

FIGURE 6.14.: À gauche, moyenne des jugements en fonction de l'ordre des compléments verbaux ; au centre, moyenne des jugements en fonction du statut du SP ; à droite, moyenne des jugements en fonction de l'interaction entre ordre et statut du SP.

de l'effet du statut du référent du SP sur les préférences des locuteurs. Pour prouver que l'hypothèse A est vraie, il faudrait mettre en place une expérience permettant de vérifier la non-existence de cet effet. En ce qui concerne l'hypothèse B, on peut supposer que le questionnaire n'est pas bien conçu pour mettre en évidence l'effet du statut du référent. Par exemple, il est possible que l'effet soit très faible et que le *design* de notre expérience soit trop "grossier" pour arriver à le faire émerger. De plus, les distracteurs utilisés pour cette expérience concernent des cas d'inversion du sujet pour lesquels les jugements peuvent être très nets. Les jugements marqués portés sur les distracteurs pourraient aussi avoir écrasé les jugements plus nuancés attendus pour le statut informationnel du référent.

Les observations sur corpus, ainsi que le questionnaire psycholinguistique, n'ont pas permis de statuer sur le rôle de l'opposition *donné-nouveau* dans l'ordonnancement des compléments postverbaux. Il nous semble, cependant, que l'effet de cette contrainte préférentielle, s'il existe, doit être relativement faible étant donné qu'il n'a pas été repérable au moyen des outils que nous avons utilisés.

6.4. Bilan

Nous avons mené un travail sur une problématique peu étudiée en français : l'ordre des compléments postverbaux. Nous nous sommes appuyée sur deux ensembles de données extraites de corpus écrits et oraux, ainsi que sur deux questionnaires.

Après avoir mené une étude préliminaire qui a permis de mieux cerner la notion de poids en français et de mettre en lumière l'importance du lemme verbal dans le choix de l'ordre des compléments postverbaux, nous avons constitué une table de données regroupant des extraits de quatre corpus et nous en avons proposé une modélisation. Nous avons ensuite discuté plus précisément du rôle de trois facteurs : le lemme verbal, le caractère animé et le statut du référent. D'après la revue de littérature que

nous avons faite dans le chapitre 5, les contraintes liées au caractère défini, animé et au statut des référents sont pertinentes dans d'autres langues. Les résultats obtenus suggèrent que ces contraintes ne sont pas actives en français, ce qui constituerait une singularité de cette langue.

6.4.1. Ordre des compléments par défaut

Selon Blinkenberg (1928, p. 179), « *c'est normalement le complément direct qui précède le [complément indirect]* », ce qui revient à dire que l'ordre par défaut, en termes catégoriels, est SN SP. Les données dont nous disposons vont dans ce sens. Premièrement, la proportion d'ordre SN SP est d'environ 70% dans *TF*. Deuxièmement, si l'on considère uniquement les phrases dans lesquelles les deux constituants ont la même longueur, cet ordre s'observe dans 168 phrases sur 215, c'est-à-dire à plus de 78%. Troisièmement, dans les deux questionnaires que nous avons fait passer, l'ordre SN SP a été significativement mieux noté que l'ordre inverse. L'ensemble de ces éléments nous amène à une conclusion similaire à celle de Blinkenberg : l'ordre par défaut en français est SN SP pour deux compléments sous-catégorisés par le verbe. Néanmoins, comme l'avait noté Blinkenberg (1928, p. 180) et comme le montre l'étude de l'ordre dans la table préliminaire, il existe un contexte dans lequel l'ordre par défaut n'est quasiment jamais respecté : il s'agit du cas de figure où l'objet est réalisé comme une infinitive ou une subordonnée. Enfin, l'idée d'un ordre par défaut peut être remise en cause par les préférences verbales que nous avons mises à jour. En effet, on pourrait considérer qu'il n'existe pas d'ordre par défaut pour les compléments de façon générale, mais un ordre par défaut pour chaque verbe.

6.4.2. La contrainte de poids

Notre travail a permis de mieux caractériser la notion de poids syntaxique par rapport à l'ordonnancement des compléments en français. Nous avons montré que l'estimation du poids en termes de syllabes ne constitue pas une aussi bonne mesure que le nombre de mots. Cela laisse penser que le poids doit être défini, au moins partiellement, par rapport à la complexité syntaxique des constituants. L'observation des mesures en nombre de noeuds syntaxiques et de noeuds syntagmatiques n'offre pas de meilleure approximation de la notion de poids. Nous en concluons que le nombre de mots est une bonne estimation du poids des constituants, dans la mesure où elle rend compte à la fois de la longueur et de la complexité.

Cette idée doit être confirmée à l'aide d'une expérience qui mettrait en jeu des constituants de longueur équivalente mais de complexité différente, comme l'ont fait Wasow & Arnold (2003).

6.4.3. Perspectives de recherche

Le travail présenté dans ce chapitre correspond à une étude exploratoire sur un sujet très peu étudié. Les perspectives de poursuites de ce travail sont donc multiples. Nous

en donnons trois qui nous semblent particulièrement intéressantes.

6.4.3.1. Autres contraintes

La modélisation de la table *TF* doit être améliorée. Nous avons déjà ouvert une piste de recherche en observant l'effet du statut donné ou nouveau des référents dans le discours. Il faudrait poursuivre dans cette direction en observant d'autres phénomènes liés à l'organisation du discours. Par exemple, il pourrait être intéressant d'étudier l'ordre des compléments dans la perspective de la cohérence textuelle. Green (1980) a mis en évidence la fonction connective (*connective function*) de l'inversion du sujet en anglais. Le syntagme antéposé au verbe, dans une cas d'inversion du sujet, peut servir à créer une cohérence discursive, c'est-à-dire à faire le lien entre le contexte gauche et le contexte droit. On pourrait émettre l'hypothèse que dans le cas des compléments postverbaux, le constituant produit le plus à droite a une fonction connective par rapport aux éléments se trouvant dans le contexte droit. D'autres aspects relatifs à l'organisation du discours pourraient être étudiés, tels que la pertinence ou l'importance de l'information véhiculée par les deux constituants. Au niveau sémantique, il faudrait développer l'étude du lien sémantique entre le verbe et l'un de ses compléments. Dans notre travail, cet aspect est pris en compte par la variable `SPfige` et semble être pertinent pour expliquer l'occurrence d'ordre SP SN. Cependant, peu de données sont concernées par cette variable dans la table *TF*, ce qui la rend non significative dans la modélisation. Le poursuite de ce travail pourrait se faire par une étude plus systématique des cas où la séquence V SP présente un caractère figé. On pourrait également envisager d'examiner d'autres types de séquences en utilisant des tests d'implication, comme dans Hawkins (2000).

6.4.3.2. Sous-problèmes

Comme nous l'avons exposé en début de chapitre, les données étudiées présentent une diversité qui pourrait expliquer la non-significativité des contraintes relatives au caractère défini ou animé des référents sur l'ordre des compléments. Par exemple, dans les expressions figées non-connexes, le SP a tendance à apparaître directement après le verbe. Or, dans ces cas, le SP est indéfini et inanimé, ce qui va à l'encontre des tendances générales selon lesquelles les constituants définis et animés tendent à précéder les indéfinis et inanimés. L'une des pistes permettant de faire émerger le rôle des contraintes non-significatives dans notre travail pourrait donc être de diviser le phénomène étudié en sous-problèmes. Par exemple, on pourrait s'intéresser plus spécifiquement à l'ordre des compléments pour les verbes sous-catégorisant un SP datif ou un SP locatif.

6.4.3.3. Biais verbaux

Il conviendrait d'approfondir les recherches concernant les biais verbaux. Nous estimons que les préférences verbales doivent être comprises comme une combinaison de facteurs intervenant à différents niveaux. La poursuite du travail sur les préférences

verbales doit se faire selon deux axes. D'un côté, il faut analyser les différentes strates intervenant dans le façonnement des préférences, afin de comprendre comment ces dernières sont formées, comme nous l'avons évoqué dans la section 6.3.1. De l'autre côté, il faut décrire de façon plus systématique les biais de chaque verbe. Pour cela, il faut recueillir plus de données relatives aux lemmes verbaux mal représentés dans nos corpus et utiliser des méthodes expérimentales pour tenter de confirmer l'existence de ces préférences dans le savoir langagier des locuteurs.

6.4.3.4. Contraintes relatives au traitement cognitif

En plus de l'ensemble des aspects que nous avons abordé, l'ordre des constituants est affecté par des contraintes directement en lien avec des phénomènes cognitifs. Ces aspects n'ont pas été étudiés ici, mais ils représentent une autre perspective de recherche. Nous présentons deux de ces phénomènes : les ambiguïtés de rattachement et l'effet de persistance syntaxique.

Eviter l'ambiguïté Certains auteurs affirment que l'ordre relatif de deux constituants peut être choisi afin d'éviter une ambiguïté de rattachement dans l'analyse de la phrase par l'interlocuteur.

Blinkenberg (1928) mentionne déjà cet idée à propos du français. Il explique que dans la phrase (25-b), le rattachement du SP *à la France* est ambigu, alors que la phrase (25-a) ne présente aucune ambiguïté.

(25) a. *Il sut parler à la France le langage qui convenait*
b. *Il sut parler le langage qui convenait à la France*
(tirées de Blinkenberg, 1928, p. 180)

Une ambiguïté de rattachement est une charge de traitement significative pour l'interlocuteur, dans la mesure où il doit produire plusieurs analyses simultanées. Or, la flexibilité de l'ordre des mots permet de produire des séquences non-ambiguës et donc de faciliter la compréhension de l'interlocuteur. Des travaux ont été menés en anglais afin de déterminer si l'ordre est effectivement utilisé pour éviter les ambiguïtés potentielles.

Wasow & Arnold (2003) ont étudié l'effet de l'ambiguïté dans l'alternance dative, à partir d'un questionnaire. Ce questionnaire contenait des paires de phrases, dont l'une était la contrapartie de l'autre d'après l'alternance dative. La moitié des paires comportait une phrase avec ambiguïté, l'autre moitié n'en contenait pas. Les participants devaient dire quelle phrase de la paire semblait plus naturelle. La construction à double objet a été significativement préférée quand la contrepartie prépositionnelle était ambigüe. Cela semble indiquer que, dans une tâche métalinguistique telle que le jugement d'acceptabilité, les locuteurs préfèrent éviter l'ambiguïté pour faciliter leur compréhension, dans les cas d'alternance dative. Partant de cette idée, les auteurs ont cherché à savoir si la production des locuteurs allait dans le même sens et avait tendance à faciliter la compréhension de l'interlocuteur. Pour cela, ils ont

conçu une expérience dans laquelle deux participants étaient impliqués, un locuteur et un interlocuteur. Le locuteur voyait une phrase contenant les informations à transmettre à l'interlocuteur (=stimuli). L'interlocuteur posait une question qui lui était imposée et qui contenait un verbe impliquant l'alternance dative (du type *give*). Le locuteur répondait à cette question en essayant de transmettre l'ensemble des informations et en réutilisant la plupart du temps le verbe à alternance dative. La moitié des stimuli présentée au locuteur était potentiellement ambigüe, dans le sens où la phrase produite en réponse à la question de l'interlocuteur pouvait contenir une ambiguïté de rattachement, si la construction prépositionnelle était choisie. Les auteurs ont classé les réponses selon que le bénéficiaire apparaissait en premier (construction à double-objet) ou que le thème apparaissait en premier (construction prépositionnelle). Si l'ambiguïté avait une influence dans le choix de la production des phrases, on s'attendrait à ce que les stimuli ambigus déclenchent des réponses contenant le bénéficiaire en premier, ce qui permet d'éviter l'ambiguïté. Les résultats de cette expérience montrent que les stimuli potentiellement ambigus déclenchent significativement moins de constructions présentant un bénéficiaire adjacent au verbe, ce qui va à l'encontre de ce qui était attendu. On observe des effets significatifs du biais lexical, de la longueur, ainsi que d'un facteur 'ambiguïté inverse', c'est-à-dire que les constructions à thème adjacent au verbe sont significativement plus fréquentes dans le cas de stimuli potentiellement ambigus. Les auteurs avancent, comme hypothèse explicative, que la présence de la préposition *to* dans le stimuli pousse les locuteurs à utiliser cette même préposition pour le bénéficiaire du verbe [36]. En conclusion, les locuteurs n'utilisent pas les possibilités offertes par l'ordonnancement des mots pour éviter les ambiguïtés d'attachement, alors que les interlocuteurs semblent préférer l'ordre qui lève l'ambiguïté.

Un travail à partir des données de corpus ainsi que dans une perspective expérimentale mériterait d'être mené pour le français et permettrait de savoir si les stratégies des locuteurs et interlocuteurs en français sont similaires à celles rencontrées pour l'anglais.

Effet de persistance syntaxique L'effet de persistance syntaxique peut se définir comme la tendance des locuteurs à réutiliser une construction syntaxique récemment utilisée ou entendue [37]. Cet effet peut affecter le choix d'un ordre des mots. Par exemple, si un locuteur vient d'entendre l'ordre V SN SP, il tendra à réutiliser cet ordre. Cette idée a été étudiée d'un point de vue expérimental notamment par Bock (1986) et Branigan *et al.* (1999), ainsi qu'à partir de données de corpus (Estival, 1985; Gries, 2005; Szmrecsanyi, 2005, parmi d'autres).

Bock (1986) a étudié l'effet de persistance dans le choix de la voix active ou passive et dans l'alternance dative. Dans les trois expériences mises en place, les sujets

36. Il s'agirait d'une forme de *priming*, ou persistance syntaxique, comme nous le verrons dans la section suivante.

37. L'effet de persistance ou *priming* n'est pas restreint à la dimension syntaxique. Il existe au niveau sémantique, lexical, morphologique…

devaient répéter une phrase entendue (phrase d'amorçage), puis décrire une image sans rapport avec ladite phrase. Bock a observé un effet significatif de la persistance syntaxique : les locuteurs ayant eu à répéter une phrase contenant la construction à double objet penchent vers l'utilisation de cette même construction pour décrire l'image, alors qu'aucun élément lexical ou sémantique n'établit un lien entre la phrase d'amorçage et la description de l'image. Bock (1986) en conclut que c'est bien la construction syntaxique qui est l'objet d'une persistance.

Szmrecsanyi (2005) étudie le choix effectué par les locuteurs pour trois constructions dans des corpus oraux de l'anglais. Parmi les trois constructions étudiées, cet auteur observe la construction Verbe - Particule en s'appuyant sur les données de *The Freiburg English Dialect Corpus*. Szmrecsanyi prend en compte les facteurs généraux connus pour influencer le choix entre la séquence *V Particule OD* et la séquence *V OD Particule* : le caractère défini, le caractère nouveau, la longueur, la complexité syntaxique, le caractère non-compositionnel de la séquence, la présence d'un SP directionnel après la séquence considérée et la préférence de chaque lemme verbal pour un ordre. La persistance syntaxique et les contextes la favorisant ont été évalués au moyen de 4 variables :

- ordre attesté dans la précédente occurrence d'une construction Verbe - Particule ;
- distance entre l'occurrence étudiée et la précédente occurrence d'une construction Verbe - Particule ;
- lemme verbal intervenant dans la précédente occurrence identique ou différent de celui de l'occurrence étudiée ;
- longueur de la phrase dans laquelle l'occurrence étudiée apparaît (façon d'approximer la complexité de la phrase produite).

En s'appuyant sur une modélisation statistique, Szmrecsanyi montre que la prise en compte des quatre variables relatives à la persistance renforce significativement le pouvoir prédictif du modèle. L'ordre précédemment produit influe sur le choix de l'ordre attesté. Plus précisément, cette influence décroît lorsque la distance entre les deux occurrences augmente ; mais elle est plus forte si le même item verbal est utilisé dans les deux occurrences. Enfin, l'influence de la construction précédemment produite dépend de la complexité de la phrase dans laquelle apparaît l'occurrence étudiée : plus la phrase est complexe, plus l'effet de persistance est important.

Dans le même ordre d'idées, Gries (2005) fournit une étude sur corpus de l'effet de persistance syntaxique dans les phénomènes de l'alternance dative et de la construction Verbe - Particule. Il montre notamment que l'effet de persistance est dépendant du verbe impliqué dans la construction. Plus précisément, il met à jour le fait que certains lemmes verbaux sont plus résistants aux effets de persistance que d'autres. Enfin, grâce à des expériences où les locuteurs doivent répéter des phrases lues, le travail de Yi *et al.* (2012) montre que les effets de persistance syntaxique sont également conditionnés par la similarité sémantique des verbes utilisés dans les phrases d'amorçage et les phrases cibles. Les auteurs observent notamment que, dans le cas de l'alternance dative, l'effet de persistance syntaxique n'est observable que dans les cas où le verbe de la phrase d'amorçage et celui de la phrase cible ont un haut degré de similarité sémantique.

Les locuteurs ayant tendance à réutiliser les mêmes constructions, l'occurrence d'un ordre particulier peut s'expliquer par la présence du même ordre, dans un contexte gauche proche. L'amélioration de la description et de la modélisation de nos données pour le français pourrait donc se faire par l'annotation du corpus selon l'ordre attesté dans le contexte qui précède la phrase étudiée. De plus, il serait intéressant d'examiner à quel niveau intervient l'effet de persistance. En effet, étant donné la variété des verbes et des cadres de sous-catégorisation des verbes que nous étudions, on peut se demander si l'effet de persistance est valable pour l'enchaînement des deux syntagmes de façon très générale, ou bien si elle concerne les verbes ayant le même type de cadre de sous-catégorisation et donc pour des SP introduits par la même préposition, ou encore si elle ne s'observe qu'avec la même tête verbale.

Conclusion

La notion de contrainte préférentielle, son étude et sa formalisation ont été au coeur de cette thèse. Comme nous l'avons exposé dans le chapitre 1, nous sommes partie de l'hypothèse générale selon laquelle la connaissance langagière des locuteurs n'est pas définie uniquement en termes de contraintes catégoriques, mais qu'elle embrasse également des préférences. Ces préférences s'observent dans la langue sous forme de contraintes préférentielles qui interviennent dans des phénomènes non catégoriques pour lesquels il existe plusieurs réalisations concurrentes, par exemple, l'alternance dative en anglais. Ces contraintes sont censées agir en production comme en compréhension et toucher à différents niveaux linguistiques tels que la prosodie, la phonologie, la syntaxe, la sémantique ou la structure informationnelle.

Pour identifier et formaliser ce type de contraintes, nous avons choisi d'utiliser la méthodologie de Bresnan *et al.* (2007) et Bresnan & Ford (2010) qui repose sur l'utilisation de données de corpus annotés et de données expérimentales. Dans le chapitre 2, le passage des observations dans les données à des propriétés générales de la langue a été envisagé grâce à l'utilisation de méthodes d'analyse de données, avec notamment la régression logistique et les modèles à effets mixtes. Grâce au premier outil statistique, on peut modéliser le comportement d'une variable binaire en fonction de plusieurs variables prédictrices. Plus exactement, cet outil permet d'évaluer la probabilité qu'une construction ou un ordre soit réalisé, étant donné un ensemble de contraintes linguistiques intégrées au modèle sous la forme de variables prédictrices. Le deuxième outil sert, grâce à l'introduction d'effets aléatoires, à prendre en compte la structuration des données. Dans le cas de la modélisation de phénomènes linguistiques, il nous a permis notamment de prendre en compte les idiosyncrasies liées au lexique. Cette méthodologie a été déployée pour deux phénomènes touchant à l'ordre des mots en français : la position de l'adjectif épithète et l'ordre des compléments postverbaux.

La position de l'adjectif épithète

Notre travail repose sur l'hypothèse explicitée dans le chapitre 3 et selon laquelle il existe en français une possibilité générale d'alternance de position pour la catégorie de l'adjectif. Cette alternance est plus ou moins forte selon les items adjectivaux

considérés. Par exemple, les adjectifs évaluatifs ont la possibilité d'apparaître en antéposition comme en postposition, tandis que les adjectifs de nationalité sont très majoritairement postposés et leur antéposition est généralement autorisée par la présence d'un adverbe tel que *très* qui assouplit les préférences lexicales.

Le travail de modélisation présenté dans le chapitre 4 a permis d'évaluer l'importance des caractéristiques lexico-sémantiques. Premièrement, les caractéristiques lexicales formelles des lemmes adjectivaux tendent à converger vers une position, comme cela est illustré en 6.22. De plus, nous avons montré que les propriétés lexicales peuvent imposer des préférences contradictoires pour un même item : les adjectifs préfixés en *in-* ont une préférence pour la postposition en raison de leur longueur élevée en moyenne, mais la présence du préfixe tend à favoriser l'antéposition.

Antéposition	NOM	**Postposition**
court		long
fréquent		rare
simple		construit

TABLE 6.22.: Convergence de faisceaux de caractéristiques lexicales selon les positions

Deuxièmement, les classes lexico-sémantiques étudiées (intensionnel, évaluatif, indéfini, nationalité) présentent des préférences claires pour une position ou pour l'autre. Cependant, la classification systématique des items selon la sémantique pose un problème méthodologique que nous avons surmonté en proposant ce que nous avons appelé un modèle lexicalisé. Dans ce modèle, les lemmes adjectivaux sont considérés comme des effets aléatoires, ce qui permet d'envisager le problème de l'alternance de position selon les spécificités formelles et sémantiques de chaque lemme. Chaque intercept aléatoire associé à un adjectif constitue un moyen d'évaluer pour quelle position l'adjectif a une préférence. Le modèle global, valable pour les adjectifs présentant une alternance de position dans nos données, montre qu'une fois les aspects lexicaux pris en compte, la place de l'adjectif est déterminée par le nom avec lequel il se combine, ainsi que par la configuration du SAdj et du SN. Premièrement, la position de l'adjectif est influencée par le nom qu'il modifie. Plus précisément, on observe que plus une séquence ordonnée Nom - Adjectif est fréquente, plus la séquence a tendance à être produite dans cet ordre, et ce, même si l'adjectif a une préférence générale pour l'autre ordre. Cela signifie que le problème de la position de l'adjectif doit être envisagé à un niveau intermédiaire qui concerne la combinaison de deux items. Deuxièmement, la position de l'adjectif est dépendante de la configuration du SAdj : plus ce dernier est complexe et long, plus il a tendance à être postposé au nom. Enfin, certains éléments du SN favorisent l'antéposition : les déterminants possessifs, définis et démonstratifs ainsi que la présence d'autres dépendants du noms tels qu'un autre adjectif ou un SP. À partir d'un questionnaire visant à éliciter les préférences de locuteurs pour 30 phrases sélectionnées en fonction du modèle global, nous avons montré l'existence d'une corrélation entre la proportion de sujets choisis-

sant la positon antéposée et la probabilité d'antéposition calculée en corpus. Il semble que cela constitue un argument pour soutenir l'idée selon laquelle les observations faites en corpus sont en correspondance avec une forme de savoir langagier.

Du point de vue formel, la modélisation statistique que nous avons utilisée a permis de rendre compte de la position attestée de 87% des adjectifs présentant une alternance de position dans nos données. De plus, le modèle offre une classification satisfaisante des données, comme en atteste la mesure *AUC* : 0.935. Le type d'approche envisagé est adéquat pour traiter ce phénomène linguistique. Il semble valider l'idée qui a guidé ce travail, selon laquelle la position de l'adjectif est un phénomène affecté par des contraintes préférentielles, qui s'expriment au niveau lexical, au niveau intermédiaire de la combinaison du nom et de l'adjectif, et au niveau syntaxique.

L'ordre relatif des compléments postverbaux

À notre connaissance, le travail présenté dans cette thèse constitue la première étude sur corpus cherchant à évaluer les contraintes influençant l'ordre des compléments postverbaux en français.

Dans le chapitre 5, en nous appuyant sur les travaux traitant de langues autres que le français, nous avons dressé le panorama des contraintes pouvant avoir une influence sur l'ordre des dépendants du verbe. En ce qui concerne le français, les travaux de Abeillé & Godard (2006); Berrendonner (1987); Blinkenberg (1928) ont mentionné le rôle de la longueur et de la complexité des constituants. Ceux de Schmitt (1987a,b) ont mis en lumière l'influence de la sémantique d'un certain nombre de verbes.

Dans le chapitre 6, nous avons proposé une étude du phénomène à partir de données de corpus extraites de quatre corpus (*French Treebank*, Est-Républicain, ESTER et CORAL-ROM). Premièrement, la notion de poids grammatical a été examinée à la lumière de ces données. En comparant les mesures de longueur (nombre de syllabes), les mesures de complexité syntaxique (nombre de noeuds syntaxiques et syntagmatiques) et les mesures en nombre de mots, nous avons constaté que la longueur seule n'est pas un indicateur suffisant pour rendre compte des phénomènes d'ordre des compléments. Le nombre de mots semble être une bonne mesure du poids grammatical puisqu'il constitue une mesure mixte qui prend en compte à la fois la longueur et la complexité. Sa capacité de prédiction étant au moins aussi élevée que celle de mesures de complexité syntaxique, plus coûteuses à obtenir, nous considérons que le nombre de mots est une bonne approximation de la notion de poids.

La modélisation statistique des données de corpus a permis de mettre en évidence l'effet du poids grammatical, et plus précisément de la différence de poids entre les deux constituants. Nous avons également montré que le lemme verbal, désambiguïsé à l'aide de catégories sémantiques annotées en contexte, avait une influence sur le phénomène. Techniquement, le lemme verbal associé à sa classe sémantique a été introduit dans le modèle comme effet aléatoire, permettant ainsi de considérer le phénomène d'ordre des compléments en tenant compte des spécificités individuelles liées à chaque tête verbale. De cette façon, nous avons pu évaluer les préférences de chaque verbe pour une position ou pour l'autre. Nous concevons les biais ver-

baux comme le reflet d'un éventail de tendances s'exprimant à différents niveaux et s'imbriquant pour former des préférences spécifiques à chaque verbe. Nous avons fourni des pistes d'analyse permettant d'apporter une explication pour une partie de ces biais verbaux. Premièrement, en suivant les travaux de Stallings *et al.* (1998), nous avons formulé l'hypothèse selon laquelle les verbes présentant la possibilité de réaliser leur objet direct sous la forme d'une subordonnée ou d'une infinitive sont disposés à être séparés de leur objet direct par un SP et favorisent donc l'ordre SP SN. D'après le modèle proposé, les biais verbaux associés aux verbes de communication − verbes qui autorisent la réalisation d'une subordonnée et/ou d'une infinitive en plus de la réalisation nominale de leur objet direct − vont dans le sens de notre hypothèse, dans la mesure où ils sont en faveur de l'ordre SP SN. Deuxièmement, à la suite de Schmitt (1987a,b), nous avons mis en évidence le rôle de la sémantique d'une classe de verbes exprimant une relation entre un patient et un état résultant [38]. Nous avons formulé l'hypothèse selon laquelle la relation établie par le verbe entre ses deux arguments a un impact sur l'ordre des constituants réalisant les arguments du verbe. Plus précisément, nous avons postulé la règle suivante : soit β et γ les deux arguments du verbe et R(β, γ) la relation qu'instaure le prédicat entre ses deux arguments, si β est affecté, au sens de Tsunoda (1985), par R(β, γ) et que γ correspond à l'état résultant de β, alors les arguments ont tendance à apparaître linéairement dans l'ordre β γ. Cette règle rend compte d'une contrainte préférentielle de nature sémantique qui s'exprime par l'intermédiaire de préférences lexicales. Les biais verbaux associés à chacun des verbes appartenant à cette classe dans le modèle sont en adéquation avec notre hypothèse.

Les observations faites sur le français présentent des singularités par rapport à d'autres langues, telles que l'anglais ou l'allemand, qui ont été plus largement étudiées du point de vue de l'organisation des dépendants du verbe. Dans nos données, les variables concernant le caractère pronominal, défini, animé et nouveau des constituants n'ont pas d'effet significatif sur l'ordre des compléments postverbaux du français. D'après les données expérimentales recueillies à l'aide de deux questionnaires d'élicitation de jugements d'acceptabilité, l'interaction du caractère animé du SP avec l'ordre de compléments, ainsi que l'interaction du caractère donné ou nouveau du SP avec l'ordre des compléments, ne présentent pas d'effet significatif sur les jugements d'acceptabilité des sujets interrogés. Comme nous l'avons mentionné dans le chapitre 6, le fait de ne pas observer l'effet d'une variable dans des données expérimentales ne permet pas de conclure que cet effet n'existe pas. Cependant, les résultats expérimentaux, ainsi que ceux obtenus grâce à la modélisation statistique, convergent et mettent en doute l'effet des contraintes liées au caractère animé, défini et nouveau en français. En ce qui concerne la pronominalité, le français se caractérise par la cliticisation massive des éléments pronominaux, ce qui explique probablement la non-significativité de ce facteur.

38. Les verbes considérés sont *compléter*, *convertir*, *ériger*, *transformer* et *faire* (dans la construction *faire [de* SN*] [*SN*]*).

Bilan

Après avoir été définie d'un point de vue théorique comme une dimension appartenant à la connaissance langagière des locuteurs, la notion de contrainte préférentielle a été confrontée à deux phénomènes d'ordre en français. Les analyses de données de corpus ont permis d'identifier différentes occurrences de contraintes préférentielles :

Position de l'adjectif

- la présence d'un déterminant défini, démonstratif ou possessif ;
- la présence d'autres dépendants du nom dans le SN (adjectifs et SP) ;
- la force d'association d'un nom et d'un adjectif dans une position particulière ;

Ordre des compléments postverbaux

- le poids relatif des constituants.

Nous avons également mis en lumière l'importance de la dimension lexicale dans les phénomènes d'ordre. Il est nécessaire d'envisager l'ordre des mots en fonction des items lexicaux, car chaque lemme présente des spécificités qui s'expriment à travers des préférences individuelles. Les spécificités en question relèvent de différents niveaux. En ce qui concerne les deux phénomènes étudiés, nous avons mis en évidence le rôle des caractéristiques formelles et lexico-sémantiques des adjectifs, ainsi que le rôle de la sémantique et du cadre de sous-catégorisation du verbe. Sur le plan linguistique, nous avons donc mis à jour que, parmi les facteurs les plus pertinents pour les phénomènes d'ordre étudiés, on compte la singularité lexicale. Notre étude pointe l'importance du lexique aussi bien au niveau de l'ordre entre les mots dans le groupe nominal, qu'au niveau syntagmatique dans le domaine postverbal.

D'un point de vue formel, les contraintes préférentielles et leur comportement sont captés à l'aide de modèles de *régression logistique à effets mixtes*. En ce qui concerne la position de l'adjectif, nous avons développé une méthodologie basée sur la comparaison de modèles. Cela nous a permis, pour un phénomène aussi complexe que la position de l'adjectif, d'évaluer l'impact des contraintes selon leur nature, selon qu'elles relèvent de la spécificité lexicale ou de la syntaxe.

L'un des intérêts de l'utilisation des modèles à effets mixtes dans le cadre d'un travail de modélisation de phénomène d'ordre des mots tient à la possibilité de capter deux dimensions en tension : la dimension relative au lexique et la dimension relative aux contraintes générales du système de la langue. En associant des effets aléatoires aux items du lexique, le modèle à effets mixtes tient compte de la structuration des données autour de chaque lemme, et traduit ainsi : la dimension locale et spécifique apportée par chaque item lexical. Simultanément, ce même modèle rend compte, grâce aux effets fixes, de tendances générales qui traversent les données et qui sont interprétées comme des contraintes préférentielles. Les modèles à effets mixtes semblent donc être adéquats pour capter la complexité des données de la langue et l'imbrication entre le niveau lexical et le niveau systémique, entre le particulier et le général.

La méthodologie déployée est également intéressante dans la mesure où elle permet une approche non-réductionniste de l'ordre des mots, à savoir une approche qui ne

réduit pas le problème de la linéarisation à un seul principe, comme par exemple chez Hawkins (1994). L'approche que nous adoptons permet de prendre en compte une variété de facteurs présentant divers degrés d'influence et de leur attribuer une importance relative à l'intérieur d'un modèle unifié. Enfin, la méthodologie utilisée constitue non seulement une procédure de validation des facteurs rencontrés dans la littérature, mais aussi une procédure de découverte, comme l'illustre la nécessité de prendre en compte la sémantique lexicale pour mieux comprendre la façon dont s'agencent les compléments postverbaux.

Nous tenons à mentionner que la qualité d'un travail s'appuyant sur une approche telle que celle développée dans cette thèse repose en partie sur les ressources disponibles. En effet, la validité des conclusions tirées d'une étude statistique en corpus est déterminée par la représentativité des données modélisées. Or, pour le français, on est confronté au nombre limité de corpus annotés disponibles. Produire des données annotées constitue un processus coûteux en termes de temps et de main-d'oeuvre. Nous espérons, à travers notre travail, avoir démontré l'intérêt linguistique et la nécessité matérielle du développement de corpus annotés en syntaxe, en sémantique et en discours, et ce, pour une grande diversité de genres.

Perspectives concernant les deux phénomènes étudiés

Les perspectives présentées ici reprennent celles qui ont déjà été évoquées dans les parties relatives aux deux phénomène étudiés.

Position de l'adjectif Le travail que nous avons présenté se concentre sur un ensemble de données extrait du corpus journalistique *French Treebank*. Afin de proposer un modèle valable pour l'ensemble de la langue et non pas seulement pour le genre journalistique, il conviendrait d'étendre le modèle à des données issues d'autres types de corpus. Nous pensons notamment à des textes littéraires et des transcriptions d'oral spontané. Les premières observations concernant les proportions d'antéposition dans les corpus ESTER, CORAL-ROM, ainsi que dans les données littéraires de Wilmet (1980) semblent montrer qu'il n'y a pas de disparités importantes selon les genres. On s'attendrait, cependant, à ce que le genre littéraire autorise une plus grande alternance de position, dans la mesure où la liberté laissée par le système de la langue est mise à profit par les auteurs à la recherche d'effets de style. Dans le cas des données d'oral, l'idée serait que les adjectifs présentent une alternance plus faible qu'à l'écrit. En effet, dans le cas de l'oral spontané, on peut supposer que la place des adjectifs soit très largement fixée par les préférences lexicales relatives à chaque adjectif.

De plus, nous avons tenté d'établir une corrélation entre les probabilités estimées en corpus et les préférences des locuteurs, en élicitant les préférences de sujets par le biais d'un questionnaire en ligne. Ce type de corrélation constitue un argument permettant de renforcer l'idée selon laquelle les résultats obtenus par les méthodes d'analyse de données en corpus sont en correspondance avec une forme de savoir langagier. La méthode utilisée, à savoir l'élicitation de préférences, pourrait être améliorée, notamment en utilisant des phrases présentant un style moins soutenu que

les phrases du journal *Le Monde*, en contrôlant mieux les conditions expérimentales et, peut-être également, en augmentant le nombre de participants. Il serait également judicieux d'utiliser d'autres tâches pour chercher des corrélations avec les probabilités calculées en corpus. On pourrait, par exemple, imaginer une tâche en compréhension reposant sur la méthode d'auto-présentation segmentée (*self-paced reading*).

Ordre des compléments postverbaux En tant qu'étude exploratoire, le travail que nous avons proposé ouvre la voie à plusieurs perspectives.

Premièrement, il est nécessaire de chercher de nouveaux facteurs pouvant expliquer l'ordre des compléments. Ainsi, il serait intéressant de prendre en compte des contraintes relatives à l'organisation du discours, ainsi que les rôles sémantiques des arguments des verbes.

Deuxièmement, l'étude proposée s'attache à dégager des contraintes préférentielles générales expliquant l'ordre de compléments sous-catégorisés par un ensemble de verbes hétérogènes. Il est possible que notre problématique mérite d'être subdivisée en sous-problèmes relatifs à des groupes de verbes présentant un comportement plus homogène. De cette façon, il pourrait apparaître que l'effet des contraintes concernant le caractère défini ou animé des constituants est significatif pour une sous-partie du problème général traité.

Troisièmement, nous envisageons d'approfondir nos recherches sur les biais verbaux. Il s'agit notamment d'investiguer les dimensions syntaxiques et sémantiques qui permettent de définir les préférences lexicales observées, en développant les pistes d'analyse ébauchées. Il sera également nécessaire de confirmer l'existence de tels biais par l'utilisation de méthodes expérimentales. Nous pensons en particulier à mettre en place un protocole équivalent à celui de Tily *et al.* (2008), qui, en observant les mouvements oculaires pendant une tâche de compréhension, a permis de montrer l'existence de préférences liées aux verbes ditransitifs de l'anglais qui s'expriment sous la forme d'attentes syntaxiques.

Enfin, afin d'avoir une image générale du problème de l'ordre des constituants postverbaux, il serait important de prendre en compte des contraintes relatives au traitement cognitif, notamment l'influence des ambiguïtés de rattachement du SP et les effets de persistance syntaxique. En ce qui concerne ce dernier facteur, il serait intéressant d'examiner à quel niveau intervient l'effet de persistance : soit au niveau général de la combinaison des SN et des SP, soit au niveau intermédiaire de verbes partageant les mêmes types de cadres de sous-catégorisation ou d'arguments sémantiques, soit au niveau lexical, pour les constructions partageant la même tête verbale.

Perspectives générales de recherche

Sur le plan linguistique, l'étude des deux phénomènes du français a montré l'importance de la dimension lexicale. D'après ce constat, on peut formuler une hypothèse de travail selon laquelle tout phénomène de linéarisation est en partie lié à la singularité des items lexicaux qu'il met en jeu. Par exemple, le problème de

l'inversion du sujet en français pourrait être envisagé dans cette perspective. De plus, le travail sur l'ordre des compléments révèle que le français semble se distinguer de l'anglais ou de l'allemand en ce qui concerne l'effet de contraintes relatives au caractère animé, défini et nouveau des constituants. Cependant, ce constat repose sur des études menées, dans chacune des trois langues, sur des phénomènes qui ne sont pas identiques. Il faudrait mettre en place un véritable travail de comparaison dans lequel les langues seraient envisagées à travers un phénomène comparable et à l'aide de données équivalentes (données expérimentales ou de corpus). De cette façon seulement, on disposerait d'arguments solides permettant la comparaison de l'existence et de l'influence des contraintes précédemment mentionnées.

Sur le plan théorique, nous avons défendu l'idée selon laquelle les contraintes préférentielles font partie de la syntaxe. Cependant, nous n'avons pas abordé la question relative à l'intégration de ces contraintes dans un modèle grammatical unifié, car cela dépassait les objectifs de notre thèse. En effet, la construction d'un cadre grammatical permettant de rendre compte à la fois des structures fondamentales des langues et des préférences n'est pas triviale et pose un certain nombre de problèmes. Premièrement, il faut disposer d'un modèle permettant de rendre compte de phénomènes de gradience. Selon les termes de Pullum & Scholz (2001), ce sont les cadres *Model-Theoretic Syntax* qui sont les mieux outillés pour décrire ces phénomènes. Dans ce type de modèle, une structure bien formée est une structure satisfaisant l'ensemble des contraintes qui définissent la grammaire. Schématiquement, en autorisant la violation de certaines contraintes, on peut rendre compte de différents degrés de bonne formation. Deuxièmement, il est nécessaire de s'interroger sur le rapport entre la notion de contrainte telle qu'elle est utilisée dans ce type de cadres et la notion de contraintes préférentielles telle que nous l'avons développée. Peut-on transférer les facteurs que nous avons étudiés sous la forme de contraintes à satisfaire dans un modèle grammatical ? Certains modèles intègrent déjà une forme de contrainte préférentielle. C'est le cas de la Théorie Stochastique de l'Optimalité (Boersma, 1998, 2000; Boersma & Hayes, 2001). Comme le montrent Bresnan *et al.* (2001), ce cadre permet de rendre compte de l'effet non catégorique de la contrainte de personne sur le choix de la voix passive en anglais [39]. Cependant, Jäger & Rosenbach (2006) ont démontré que la Théorie Stochastique de l'Optimalité ne rend pas compte des effets de cumulativité (*ganging-up cumulativity*), à savoir que, lorsque plusieurs contraintes de faible importance se combinent, leur importance respective s'additionne et leurs effets sont cumulés. Le modèle de la Théorie Linéaire de l'Optimalité (Keller, 2006) dépasse cette limite en attribuant des poids aux contraintes. Cependant, ce modèle s'appuie sur la violation de contraintes pour rendre compte des phénomènes de gradience. Or, dans les modèles statistiques que nous avons présentés, la satisfaction des contraintes est aussi pertinente que leur violation pour rendre compte des préférences. Se pose alors la question de savoir si la modélisation que nous avons proposée est traduisible en termes de violation de contraintes uniquement. Sur ce point, les Gram-

39. Les données relatives à l'interaction de la personne et de la voix en anglais sont présentées dans la section 1.1.2.1 du chapitre 1.

maires de Propriétés (Blache, 2005) présentent l'avantage de permettre de prendre en compte à la fois les contraintes violées et les contraintes satisfaites, comme cela est montré dans Blache *et al.* (2006). Troisièmement, il faut élaborer le traitement de la linéarité dans les cadres formels. En effet, les règles de précédence linéaires sont généralement réductionnistes dans la mesure où les principes d'ordre sont gérés catégoriellement (par exemple, Déterminant précède Nom), fonctionnellement (par exemple, Tête précède Non-Tête) ou grâce à des traits (par exemple, Léger précède Non-Léger). Certains raffinements ont été proposés : Abeillé & Godard (2000) proposent des règles de précédence linéaire mixtes qui combinent par exemple des traits et des fonctions. Étant donné la complexité des modèles statistiques proposés, notamment pour rendre compte de la position de l'adjectif, il semble que les règles de précédence linéaire doivent être enrichies.

En termes d'analyse statistique des données, nous avons utilisé des modèles à effets mixtes, en limitant l'introduction d'effets aléatoires sur l'intercept général du modèle. De cette façon, nous avons notamment capté les spécificités relatives aux items lexicaux par rapport aux phénomènes étudiés. On pourrait envisager l'utilisation d'effets aléatoires sur les coefficients de pente associés aux variables prédictrices. Par exemple, il est possible que certains verbes soient plus sensibles aux effets du poids grammatical que d'autres. Dans ce cas, un effet aléatoire par item lexical sur la pente associée à la mesure de poids permettrait de rendre compte de comportements différents des lemmes verbaux en ce qui concerne le poids grammatical. Pour observer ces phénomènes, il faudrait un plus grand nombre d'observations par verbe. Ce type de modèles plus complexes constituerait un outil permettant de rendre compte de façon plus précise de l'articulation entre lexique et règles générales.

Enfin, notre travail pose le problème du rapport entre modélisation à partir de données de corpus et connaissance langagière des locuteurs : dans quelle mesure les observations dégagées sur nos données sont-elles en lien avec la connaissance des locuteurs ? D'un point de vue théorique, le principal lien que nous voyons, se situe dans l'idée selon laquelle la connaissance langagière est façonnée par l'usage. Ainsi, les occurrences de phénomènes et leur fréquence agissent sur la représentation que les locuteurs ont de ces phénomènes. En étudiant la fréquence observée en corpus, on tente de capter la connaissance liée à l'usage de la langue. Ces hypothèses de travail doivent être étayées au moyen de preuves empiriques. Pour cela, il serait nécessaire de mettre en parallèle des modèles de corpus et des données issues de méthodes expérimentales. En montrant qu'il existe une correspondance entre les résultats obtenus en corpus et ceux obtenus au moyen d'expériences, comme l'ont fait par exemple Bresnan & Ford (2010), Tily *et al.* (2008), Bresnan (2007b), nous aurons des arguments forts pour affirmer que les connaissances induites sur corpus ont une réalité chez les locuteurs.

Annexe A

Questionnaire portant sur les préférences de position de l'adjectif épithète

Expérience de linguistique

Instructions

Vous allez lire successivement deux phrases qui ne se distinguent que par un seul segment. Ce segment sera présenté en bleu dans les deux phrases. Par exemple :

- *Henri Guitton a joué* **un rôle important** *dans la modernisation de l'enseignement de l'économie en France*
- *Henri Guitton a joué* **un important rôle** *dans la modernisation de l'enseignement de l'économie en France*

L'expérience porte sur les deux segments en bleu dans le contexte de la phrase. Parmi ces deux options, vous devez choisir celle que vous utiliseriez si vous deviez écrire cette phrase vous-même. Pour marquer votre préférence, vous devrez cocher la case située à gauche de la phrase choisie.

Vous devez essayer de sélectionner la séquence qui a votre préférence de la façon la plus spontanée possible : essayez de suivre votre intuition plutôt que de tenter de vous rappeler ce que vous croyez être du "bon français".

Pour chaque paire de phrases ainsi proposée, nous vous demandons de lire intégralement les deux options avant de sélectionner celle qui a votre préférence.

- Si vous le souhaitez, vous pouvez lire les phrases à haute voix avant de répondre.
- Les phrases de ce questionnaire sont inspirées de phrases rencontrées dans des journaux tels que *Le Monde*. Elles traitent pour la plupart de sujets politiques et économiques. Ne tenez pas compte de l'aspect stylistique de ces phrases. Ne

faites confiance qu'à votre réaction première.
– L'objectif de ce questionnaire est de recueillir vos préférences individuelles. Il est donc impératif de le remplir seul.

L'expérience compte 30 paires de phrases.

Attention : ne tentez pas de revenir en arrière avec la flèche de votre navigateur ou de rafraîchir la page durant l'expérience. Cela fausserait vos résultats et ces derniers ne pourraient pas être pris en compte.

Merci pour votre collaboration !

Phrases du questionnaire

Les phrases sont présentées avec la probabilité d'antéposition estimée par le Modèle Global et organisées par ordre de probabilité croissante.

Phrase 1 $P(\texttt{position} = 1|X, L_i) = 0.0019799593$, ordre attesté = (1-b)

(1) a. *Depuis quelques mois, ce dernier est d'ailleurs resté très en retrait, laissant son successeur, qui devrait être secondé par M. Alain Obadia, sur le devant de la scène.* **À l'actuelle heure,** *seule l'arrivée de trois nouveaux venus au sein du bureau confédéral est acquise.*

b. *Depuis quelques mois, ce dernier est d'ailleurs resté très en retrait, laissant son successeur, qui devrait être secondé par M. Alain Obadia, sur le devant de la scène.* **À l'heure actuelle,** *seule l'arrivée de trois nouveaux venus au sein du bureau confédéral est acquise.*

Phrase 2 $P(\texttt{position} = 1|X, L_i) = 0.033898754$, ordre attesté = (2-a)
(SN ambigu, phrase écartée des résultats.)

(2) a. *"Vous nous étranglez avec vos taux d'intérêt élevés" lancent* **les responsables britanniques** *à ceux de la Bundesbank*

b. *"Vous nous étranglez avec vos taux d'intérêt élevés" lancent* **les britanniques responsables** *à ceux de la Bundesbank*

Phrase 3 $P(\texttt{position} = 1|X, L_i) = 0.047597295$, ordre attesté = (3-a)

(3) a. *Un investisseur y trouvera un arsenal bien garni de services aux entreprises,* **des équipements modernes en télécommunications,** *une palette d'avantages financiers et surtout fiscaux (depuis la loi Pons de 1986) particulièrement attrayants.*

b. *Un investisseur y trouvera un arsenal bien garni de services aux entreprises,* **des modernes équipements en télécommunications,** *une palette d'avantages financiers et surtout fiscaux (depuis la loi Pons de 1986) particulièrement attrayants.*

Phrase 4 $P(\texttt{position} = 1|X, L_i) = 0.147929911$, ordre attesté = (4-b)

(4) a. *Encore faudrait-il que, pour faire passer la pilule des réformes nécessaires - et que beaucoup d'Italiens, en dépit de leur enthousiasme, risquent, une fois au pied du mur, de trouver plus amère que prévu, - qu'***un fort et surtout crédible gouvernement** *se constitue.*

b. *Encore faudrait-il que, pour faire passer la pilule des réformes nécessaires - et que beaucoup d'Italiens, en dépit de leur enthousiasme, risquent, une fois au pied du mur, de trouver plus amère que prévu, - qu'***un gouvernement fort et surtout crédible** *se constitue.*

Phrase 5 $P(\texttt{position} = 1|X, L_i) = 0.168494083$, ordre attesté = (5-a)

(5) a. *Des dispositions contraires à la loi, qui prévoyait au contraire* **une réglementation unique.**

b. *Des dispositions contraires à la loi, qui prévoyait au contraire* **une unique réglementation.**

Phrase 6 $P(\texttt{position} = 1|X, L_i) = 0.183588606$, ordre attesté = (6-b)

(6) a. *Cette solution vise à désamorcer les craintes du Congrès quant à un transfert* **de"sensibles technologies"** *dans des mains étrangères, en l'occurrence françaises.*

b. *Cette solution vise à désamorcer les craintes du Congrès quant à un transfert* **de "technologies sensibles"** *dans des mains étrangères, en l'occurrence françaises.*

Phrase 7 $P(\texttt{position} = 1|X, L_i) = 0.200049743$, ordre attesté = (7-a)

(7) a. *Une chose est certaine pour l'instant : le marché parisien est encombré de "papiers" portant* **des signatures plus ou moins solides** *tandis que banques et compagnies d'assurances, notamment publiques, supportent tout le poids de placements, peut-être excellents à long terme, mais difficiles à gérer dans l'intervalle.*

b. *Une chose est certaine pour l'instant : le marché parisien est encombré de "papiers" portant* **des plus ou moins solides signatures** *tandis que banques et compagnies d'assurances, notamment publiques, supportent tout le poids de placements, peut-être excellents à long terme, mais difficiles à gérer dans l'intervalle.*

Phrase 8 $P(\texttt{position} = 1|X, L_i) = 0.2579486$, ordre attesté = (8-b)

(8) a. *L'entreprise Sécuripost arrive en effet sur* **un difficile marché** *qui stagne depuis le développement de la monétique, avec un handicap sérieux : ses coûts de main-d'oeuvre.*

b. *L'entreprise Sécuripost arrive en effet sur* **un marché difficile** *qui stagne depuis le développement de la monétique, avec un handicap sérieux : ses*

coûts de main-d'oeuvre.

Phrase 9 $P(\texttt{position} = 1|X, L_i) = 0.260762242$, ordre attesté = (9-a)

(9) a. *Il faut remonter à l'été 2002 pour retrouver* **des niveaux aussi bas,** *avant l'ascension déclenchée par la chute du mur de Berlin, qui vit flamber les taux d'intérêt allemands dans la perspective d'énormes appels de fonds pour financer la réunification des deux Allemagnes.*

b. *Il faut remonter à l'été 2002 pour retrouver* **des aussi bas niveaux,** *avant l'ascension déclenchée par la chute du mur de Berlin, qui vit flamber les taux d'intérêt allemands dans la perspective d'énormes appels de fonds pour financer la réunification des deux Allemagnes.*

Phrase 10 $P(\texttt{position} = 1|X, L_i) = 0.302384900$, ordre attesté = (10-b)

(10) a. *Encore faut-il que les germes de* **ces possibles changements** *s'incarnent en offres concrètes et qu'ils soient acceptés par l'usage.*

b. *Encore faut-il que les germes de* **ces changements possibles** *s'incarnent en offres concrètes et qu'ils soient acceptés par l'usage.*

Phrase 11 $P(\texttt{position} = 1|X, L_i) = 0.351449174$, ordre attesté = (11-a)

(11) a. *Alors que* **la parité actuelle du mark et du franc** *devient l'un des thèmes importants de la campagne électorale en France, plusieurs personnalités étrangères viennent de prendre position en faveur du maintien des cours de change actuels en Europe.*

b. *Alors que* **l'actuelle parité du mark et du franc** *devient l'un des thèmes importants de la campagne électorale en France, plusieurs personnalités étrangères viennent de prendre position en faveur du maintien des cours de change actuels en Europe.*

Phrase 12 $P(\texttt{position} = 1|X, L_i) = 0.379843693$, ordre attesté = (12-b)

(12) a. *De son côté, Alitalia, qui espère équilibrer ses comptes en 1992 après plusieurs années de pertes, va bénéficier d'un plus grand accès aux marchés de l'Est et pourra utiliser* **l'ultra-moderne aéroport de Budapest.**

b. *De son côté, Alitalia, qui espère équilibrer ses comptes en 1992 après plusieurs années de pertes, va bénéficier d'un plus grand accès aux marchés de l'Est et pourra utiliser* **l'aéroport ultra-moderne de Budapest.**

Phrase 13 $P(\texttt{position} = 1|X, L_i) = 0.410971529$, ordre attesté = (13-a)

(13) a. *Le service de la dette du Nigéria s'élèvera à 5,6 milliards de dollars en 1992. Compte tenu de* **ces remboursements très lourds** *le Nigéria continue de réclamer un allègement exceptionnel de son endettement, du type de celui dont bénéficient les pays les plus pauvres.*

b. *Le service de la dette du Nigéria s'élèvera à 5,6 milliards de dollars en*

1992. Compte tenu de **ces très lourds remboursements** *le Nigéria continue de réclamer un allègement exceptionnel de son endettement, du type de celui dont bénéficient les pays les plus pauvres.*

Phrase 14 $P(\texttt{position} = 1|X, L_i) = 0.492671301$, ordre attesté = (14-b)

(14) a. *Le patronat estime que les conditions d'un accord ne sont pas réunies et envisage de faire le point de la situation courant janvier avec les syndicats.* **A l'issue de la séance de négociations précédente,** *le 22 décembre, les syndicats avaient jugé que le prix à payer en échange de la fixation d'un taux minimum de cotisation de 6% était totalement disproportionné et avaient demandé au patronat de modifier sérieusement ses propositions.*

b. *Le patronat estime que les conditions d'un accord ne sont pas réunies et envisage de faire le point de la situation courant janvier avec les syndicats.* **A l'issue de la précédente séance de négociations,** *le 22 décembre, les syndicats avaient jugé que le prix à payer en échange de la fixation d'un taux minimum de cotisation de 6% était totalement disproportionné et avaient demandé au patronat de modifier sérieusement ses propositions.*

Phrase 15 $P(\texttt{position} = 1|X, L_i) = 0.498904395$, ordre attesté = (15-a)

(15) a. *Il se demande où il va loger sa famille de trois enfants s'il a la malchance de vivre en région parisienne, combien d'heures de trajet il subira chaque jour pour aller travailler et* **combien de fins de mois difficiles** *il devra affronter.*

b. *Il se demande où il va loger sa famille de trois enfants s'il a la malchance de vivre en région parisienne, combien d'heures de trajet il subira chaque jour pour aller travailler et* **combien de difficiles fins de mois** *il devra affronter.*

Phrase 16 $P(\texttt{position} = 1|X, L_i) = 0.503050981$, ordre attesté = (16-b)

(16) a. *Les Lorrains, grâce à la présence du fer, du bois, puis du charbon, ont développé très tôt* **une sidérurgie importante,** *ainsi qu'une multitude d'activités traditionnelles (faïencerie, cristallerie, lutherie, etc.), dont les survivances sont encore nombreuses.*

b. *Les Lorrains, grâce à la présence du fer, du bois, puis du charbon, ont développé très tôt* **une importante sidérurgie,** *ainsi qu'une multitude d'activités traditionnelles (faïencerie, cristallerie, lutherie, etc.), dont les survivances sont encore nombreuses.*

Phrase 17 $P(\texttt{position} = 1|X, L_i) = 0.508382256$, ordre attesté = (17-a)

(17) a. *Largement dues à l'aggravation du chômage, les actuelles difficultés financières du régime de l'UNEDIC révèlent au grand jour les ambiguïtés qui sourdaient déjà dans l'accord du 18 juillet quand, pour réduire un*

déficit estimé alors à 20 milliards d'euros, les partenaires sociaux avaient été amenés à **d'importantes révisions.**

b. *Largement dues à l'aggravation du chômage, les actuelles difficultés financières du régime de l'UNEDIC révèlent au grand jour les ambiguïtés qui sourdaient déjà dans l'accord du 18 juillet quand, pour réduire un déficit estimé alors à 20 milliards d'euros, les partenaires sociaux avaient été amenés à* **des révisions importantes.**

Phrase 18 $P(\texttt{position} = 1|X, L_i) = 0.564016207$, ordre attesté = (18-b)

(18) a. *Peter Sutherland possède le professionnalisme, la crédibilité et le tempérament nécessaires pour accomplir* **la réforme indispensable,** *celle-là qui devrait conduire à la création d'une organisation mondiale du commerce adaptée aux données actuelles de l'échange.*

b. *Peter Sutherland possède le professionnalisme, la crédibilité et le tempérament nécessaires pour accomplir* **l'indispensable réforme,** *celle-là qui devrait conduire à la création d'une organisation mondiale du commerce adaptée aux données actuelles de l'échange.*

Phrase 19 $P(\texttt{position} = 1|X, L_i) = 0.6161940064$, ordre attesté = (19-a)

(19) a. *Un même consommateur a des sensibilités aux marques différentes selon les marchés et selon les situations d'achat, expliquent les auteurs : le même qui choisira avec un détachement complet, au petit bonheur, ses piles électriques, sa lessive ou ses yaourts, fera preuve d'une grande circonspection pour l'acquisition d'un matelas, d'une bouteille de champagne ou d'une eau de toilette - cas où la marque constitue* **une information précieuse.**

b. *Un même consommateur a des sensibilités aux marques différentes selon les marchés et selon les situations d'achat, expliquent les auteurs : le même qui choisira avec un détachement complet, au petit bonheur, ses piles électriques, sa lessive ou ses yaourts, fera preuve d'une grande circonspection pour l'acquisition d'un matelas, d'une bouteille de champagne ou d'une eau de toilette - cas où la marque constitue* **une précieuse information.**

Phrase 20 $P(\texttt{position} = 1|X, L_i) = 0.623632522$, ordre attesté = (20-b)

(20) a. *A telle enseigne que c'est pratiquement toute l'industrie des biens d'équipement qui se trouve, peu ou prou, engagée dans* **ce processus périlleux :** *pour attirer une clientèle qui se dérobe, les vendeurs sont amenés à consentir des rabais de plus en plus importants.*

b. *A telle enseigne que c'est pratiquement toute l'industrie des biens d'équipement qui se trouve, peu ou prou, engagée dans* **ce périlleux processus :** *pour attirer une clientèle qui se dérobe, les vendeurs sont amenés à consentir des rabais de plus en plus importants.*

Phrase 21 $P(\texttt{position} = 1|X, L_i) = 0.667660259$, ordre attesté = (21-a)

(21) a. *Les dossiers de recours en annulation ont été déposés par quatre firmes américaines, les fabricants d'ordinateurs Asus et Apple, les sociétés de services GTEI et Electronic Data Systems.* **Des multiples points soulevés par les concurrents malheureux de ZDS,** *deux ont retenu l'attention des juges.*

b. *Les dossiers de recours en annulation ont été déposés par quatre firmes américaines, les fabricants d'ordinateurs Asus et Apple, les sociétés de services GTEI et Electronic Data Systems.* **Des points multiples soulevés par les concurrents malheureux de ZDS,** *deux ont retenu l'attention des juges.*

Phrase 22 $P(\texttt{position} = 1|X, L_i) = 0.727985337$, ordre attesté = (22-b)

(22) a. *Le groupe Marland Distribution, rebaptisé Kléber 55 après* **la récente démission de M. François Marland pour raisons de santé,** *va être racheté par un fonds d'investissement, a annoncé la société mercredi 23 décembre.*

b. *Le groupe Marland Distribution, rebaptisé Kléber 55 après* **la démission récente de M. François Marland pour raisons de santé,** *va être racheté par un fonds d'investissement, a annoncé la société mercredi 23 décembre.*

Phrase 23 $P(\texttt{position} = 1|X, L_i) = 0.765996543$, ordre attesté = (23-a)

(23) a. *Une fois acquise la question de l'équilibrage structurel et donc résolue celle du déficit financier par* **des mesures nouvelles,** *M. Pierre Gilson a annoncé qu'il y a peut-être quelque chose à trouver du côté des entreprises et des salariés, qui serait à aborder en son temps, y compris avec l'Etat.*

b. *Une fois acquise la question de l'équilibrage structurel et donc résolue celle du déficit financier par* **de nouvelles mesures,** *M. Pierre Gilson a annoncé qu'il y a peut-être quelque chose à trouver du côté des entreprises et des salariés, qui serait à aborder en son temps, y compris avec l'Etat.*

Phrase 24 $P(\texttt{position} = 1|X, L_i) = 0.799713265$, ordre attesté = (24-b)

(24) a. *Sur le marché obligataire, très agité, la hausse des taux a d'abord connu* **un recul vif des obligations à échéance 10 ans,** *puis, vendredi, un vif rebond à 111,12 points, beaucoup d'opérateurs soldant leur engagement à l'approche de la fin d'année.*

b. *Sur le marché obligataire, très agité, la hausse des taux a d'abord connu* **un vif recul des obligations à échéance 10 ans,** *puis, vendredi, un vif rebond à 111,12 points, beaucoup d'opérateurs soldant leur engagement à l'approche de la fin d'année.*

Phrase 25 $P(\texttt{position} = 1|X, L_i) = 0.811582634$, ordre attesté = (25-a)

(25) a. *Les réassureurs, qui souffrent de la multiplication de catastrophes naturelles* **ces dernières années,** *ont été particulièrement touchés, à l'image de la SCOR, en baisse de 15%.*
b. *Les réassureurs, qui souffrent de la multiplication de catastrophes naturelles* **ces années dernières,** *ont été particulièrement touchés, à l'image de la SCOR, en baisse de 15%.*

Phrase 26 $P(\texttt{position} = 1|X, L_i) = 0.849142579$, ordre attesté = (26-b)

(26) a. *L'Italie connaît depuis le second semestre 2008 un ralentissement progressif de son économie qui fait suite à une période faste de six années d'expansion.* **La croissance italienne faible des trois dernières années** *a entraîné une détérioration continue du climat de confiance.*
b. *L'Italie connaît depuis le second semestre 2008 un ralentissement progressif de son économie qui fait suite à une période faste de six années d'expansion.* **La faible croissance italienne des trois dernières années** *a entraîné une détérioration continue du climat de confiance.*

Phrase 27 $P(\texttt{position} = 1|X, L_i) = 0.849240501$, ordre attesté = (27-a)

(27) a. *Une fois les réformes lancées, les Occidentaux consentiraient, alors,* **un soutien technique et financier important,** *sous l'égide des organismes internationaux.*
b. *Une fois les réformes lancées, les Occidentaux consentiraient, alors,* **un important soutien technique et financier,** *sous l'égide des organismes internationaux.*

Phrase 28 $P(\texttt{position} = 1|X, L_i) = 0.952326145$, ordre attesté = (28-b)

(28) a. *Même différence en ce qui concerne la Communauté et son avenir : pour* **le directeur nouveau du GATT, favorable à une intégration poussée,** *la construction européenne ne se limite certainement pas à une affaire de commerce, à la création d'une zone de libre-échange.*
b. *Même différence en ce qui concerne la Communauté et son avenir : pour* **le nouveau directeur du GATT, favorable à une intégration poussée,** *la construction européenne ne se limite certainement pas à une affaire de commerce, à la création d'une zone de libre-échange.*

Phrase 29 $P(\texttt{position} = 1|X, L_i) = 0.9756816$, ordre attesté = (29-a)

(29) a. *C'est ce qu'on appelle,* **en bon français,** *un dilemme.*
b. *C'est ce qu'on appelle,* **en français bon,** *un dilemme.*

Phrase 30 $P(\texttt{position} = 1|X, L_i) = 0.999071221$, ordre attesté = (30-b)

(30) a. *A Bruxelles, où il fut commissaire de 1985 à 1989, chargé d'abord des affaires sociales et de l'éducation, puis responsable de la politique de concurrence, Peter Sutherland semble faire l'unanimité. Aussi est-ce avec* **une satisfaction grande** *que sa désignation à la tête du GATT y a été accueillie, comme si cet Irlandais qui a réussi était paré de toutes les qualités pour s'acquitter avec efficacité et équité de cette mission difficile.*

b. *A Bruxelles, où il fut commissaire de 1985 à 1989, chargé d'abord des affaires sociales et de l'éducation, puis responsable de la politique de concurrence, Peter Sutherland semble faire l'unanimité. Aussi est-ce avec* **une grande satisfaction** *que sa désignation à la tête du GATT y a été accueillie, comme si cet Irlandais qui a réussi était paré de toutes les qualités pour s'acquitter avec efficacité et équité de cette mission difficile.*

Annexe B

Guide d'annotation

Le caractère animé (*animacy* en anglais) est une propriété sémantique des entités [1]. On peut faire référence à ces entités grâce à différentes expressions linguistiques.

L'*animacy* est généralement conçue selon une hiérarchie ou échelle. On peut imaginer plusieurs hiérachies.

- la plus simple : humain > non-humain
- un peu plus élaborée : humain > animal > inanimé
- encore plus élaborée :

HUMAIN	>	ANIMÉ	>	INANIMÉ
humain		animal		concret
		organisation		non-concret
		machine intelligente		lieu
		véhicule		temps

Nous allons nous appuyer sur cette dernière hiérarchie pour la tâche d'annotation du corpus.

*** La chose essentielle à retenir en annotant l'*animacy*, c'est que vous n'annotez pas le mot d'un point de vue abstrait, mais **l'entité à laquelle il fait référence** dans la phrase donnée ***.
Par exemple, le mot *église* renvoie à première vue à un INANIMÉ CONCRET, mais en contexte, il peut faire référence à une institution, un groupe de personnes...

De plus, durant l'annotation, vous devez prendre en compte l'intégralité du SN pour déterminer l'*animacy*.
Par exemple, dans l'expression *la Maison Blanche*, si vous ne considérez que le nom tête (*maison*), vous serez tenter de coder CONCRET, alors que le SN entier peut renvoyer à LIEU ou ORGA selon le contexte.

1. Le guide d'annotation pour le caractère animé est une adaptation au français du guide d'annotation de Garretson (2004) pour l'anglais.

humain

Cette étiquette est utilisée pour les SN qui réfèrent à un ou plusieurs humains. Il peut s'agir de :
- nom propres (*Serge Gainsbourg*),
- de termes de parenté (*père*, *cousin*),
- de noms communs (*linguiste*, *patron*, *étudiant*)
- de pronom (*lui*, *chacun*)

Remarques :

1. l'entité à laquelle réfère le NP doit avoir les caractéristiques d'un humain, elle ne doit pas forcément exister dans le monde réel : les morts et les personnages humains de fiction sont codés HUMAIN. Par exemple, *Luke Skywalker* doit être codé HUMAIN.
2. Dans le même ordre d'idées, les entités humanoïdes telles que les dieux, les elfes et autres fantômes portent l'étiquette HUMAIN.
3. Enfin, le pronom négatif *personne*, comme dans *il n'a donné de punition à personne*, est considéré comme HUMAIN, car même s'il n'y a pas d'entité à laquelle réfère directement le mot, nous considérons qu'il a des caractéristiques d'humain (et non pas d'animal ou de chose inanimée).

organisation

Cette étiquette est utilisée pour marquer les SN qui réfère à des organisations, des institutions, des entités collectives composées d'humains (pas d'animaux). Par exemple, le mot *Microsoft* est souvent utilisé pour référer à un ensemble de personnes qui composent l'entreprise.

Le problème de cette étiquette est qu'elle est difficile à définir et par conséquent difficile à utiliser. Par exemple, dans la phrase, *Washington n'a pas envoyé de sénateurs au Congrès américain*, on peut se demander à quoi réfère exactement *Washington* :
- soit aux habitants de Washington qui devraient être représentés par les sénateurs (HUMAIN)
- soit à la ville comme entité politique (ORGA ou NON-CONC)

Les deux critères qui doivent aider à choisir l'étiquette ORGA sont :

1. il faut que la collectivité soit composée d'humains
2. il faut que plusieurs humains soient impliqués et qu'il y ait un certain degré d'identité de groupe. Par exemple, l'expresseion *deux garçons* ne doit pas être étiquetée ORGA. Pour aider à différencier groupe d'humains marqué HUMAIN et groupe d'humains marqué ORGA, voici un ensemble des propriétés qui forment une hiérarchie implicationnelle :
   +/− `officiel`

+/− `stable dans le temps`
+/− `but/voix collective`
+/− `action collective`
+/− `collectif`

Si un groupe est + `officiel`, il sera aussi + `stable dans le temps`, + `voix collective` etc...

L'idée pour l'annotation est qu'un groupe d'humains qui a la propriété + `voix collective` ou une propriété supérieure (+ `stable dans le temps`, + `officiel`) sera annoté ORGA. Ainsi, *les Stups* sera codé ORGA, mais *la foule* sera annoté HUMAIN, et non pas organisation.
Attention, dans certains emplois, un nom d'entreprise peut renvoyer à autre chose qu'à ORGA. Par exemple, dans la phrase *Microsoft a été fondé en 1980*, *Microsoft* doit être étiqueté comme NON-CONCRET car il ne renvoie pas aux gens composant l'entreprise, mais plutôt à l'entité inanimée.
De même, dans la phrase *La France a signé le traité*, *France* renvoie à une ORGA, alors que dans *Le traité a été signé en France*, *France* doit être annoté LIEU.

animal

Cette étiquette est utilisée pour les SN qui réfèrent à des animaux non-humains. Il peut s'agir de :
- nom propre (*Médor*)
- pronom (*il*)
- nom commun (*créature*, *chien*)

vehicule

Cette étiquette est utilisée pour les véhicules car même si ce ne sont pas des êtres animés au même titre que les humains ou les animaux, ils présentent une mobilité et une possibilité de mouvement qui en font une catégorie intermédiaire entre animé et concret inanimé.

machine

Cette étiquette est utilisée pour les machines "intelligentes" telles que les ordinateurs ou les robots, car de la même façon que pour les véhicules, les machines ont plus de caractéristiques communes avec des êtres animés et réfléchis qu'avec de simples objets concrets.

concret

Cette étiquette est utilisée pour les SN qui ont un référent concret. La définition que nous donnons à concret est relativement restrictive ici : n'est concret que ce qui est *prototypiquement* concret, c'est-à-dire qui peut être perçu clairement avec l'un de nos 5 sens (au moins). Voici des exemples d'objets prototypiquement concrets :

> pomme, lit, voiture, carton, tasse, terre, porte, essence, pistolet, maison, genou, couteau, échelle, viande, lune, stylo, rivière, rocher, ciseau, chemise, neige, cuillère, étoile, bâton, sucre, soleil, table, arbre, eau

Voici quelques noms qui ne sont pas prototypiquement concrets, et qui auront donc l'étiquette NON-CONCRET :

> air, atome, chromosome, courant, brouillard, fumée, molécule, protéine, voix, vent, marée, radiation...

Remarques :

1. les parties du corps sont codées CONCRET
2. la vache est codée ANIMAL, mais du boeuf, en tant que aliment, est codé CONCRET

Exemples issus de l'annotation de l'Est-Républicain :

⋆ ph 34 : *Jean-Louis Piquard vient d'ajouter un nouveau wagon à l' hôtel de la Gare, en construisant, sur le même style que le précédent, un nouvel élément hôtelier*
- *à l' hôtel de la Gare* : CONCRET plutôt que LIEU, car on parle du bâtiment comme d'un objet qui doit être modifié.

non-concret

Cette étiquette est utilisée pour un groupe de SN très large et très varié. C'est l'étiquette "par défaut" quand une entité n'est pas animée et n'est pas concrète. Par exemple, les évènements sont NON-CONC : *voyage*, *pensée*, *rire*...

Exemples issus de l'annotation de l'Est-Républicain :

⋆ ph 57 : *Si l'on raisonne par contre en termes d'agglomération, c'est-à-dire en ajoutant à chaque ville de Haute-Saône les communes de sa périphérie qui sont dans la continuité urbaine, c'est l'agglomération de Luxeuil qui venait en tête, avec près de 13.000 habitants, contre 12.000 sur Héricourt et Lure*
- *à chaque ville de Haute-Saône* : NON-CONCRET car on fait référence à l'entité abstraite que constitue la ville, en termes administratifs
- *les communes de sa périphérie qui sont dans la continuité urbaine* : NON-CONCRET (idem)

⋆ ph 141 *Jean Page, le président de la foire, avait donc le sourire hier soir en annonçant la nouvelle à la presse et en présentant dans ces grandes lignes cette septième édition qui mettra au pinacle "Le Canada et ses trappeurs" (nous y reviendrons dans notre édition de vendredi)*

– *à la presse* : NON-CONC

⋆ ph 355 *Et elle doit ce dernier miracle à une nouvelle génération*

– *à une nouvelle génération* : NON-CONC

lieu

L'étiquette LIEU est utilisée pour les SN qui font référence à un lieu en tant qu'endroit stable et lieu potentiel pour un humain. Cela signifie que *cabine téléphonique* est codé LIEU, alors que *tiroir* est annoté CONC.

Remarques :

1. Un nom de lieu n'est pas automatiquement étiqueté LIEU. Par exemple, *Paris* peut désigner un lieu, mais aussi les habitants de Paris, le conseil municipal de Paris, le gouvernement français...
2. Il faut faire attention à l'emploi des SN en contexte. Par exemple, *sa maison* sera codé LIEU dans *je suis allée dans sa maison*, mais il sera annoté CONCRET dans *sa maison a été construite en 1920* ou *sa maison est en brique*.

temps

Cette étiquette est utilisée pour marquer les SN qui font référence à une période de temps. La période peut être courte ou longue : *un moment* ou *le 18ème siècle*. L'idée est qu'il faut pouvoir mettre ce à quoi réfère le SN sur une ligne du temps, une frise chronologique.

Annexe C

Questionnaire sur le caractère animé du SP

Ce questionnaire a été réalisé par Anne Abeillé et Benoît Crabbé.

Expérience de linguistique

Age :

Langue maternelle :

Autres langues parlées :

Région où vous avez grandi :

Nous voudrions que vous vous mettiez dans la peau d'un informant qui peut dire si les phrases qu'il va lire lui semblent naturelles.

Il n'est pas question de juger si ces phrases respectent la grammaire ou la typographie scolaire, il s'agit plutôt de dire spontanément si ces phrases seraient appropriées dans un contexte d'usage classique de la langue.

A chaque fois, vous avez le choix entre deux continuations. On vous demande de mettre une note indépendamment à chaque continuation en utilisant une échelle à 5 valeurs (1 note la moins bonne, 5 note la meilleure) :

1	**2**	**3**	**4**	**5**
☐	☐	☐	☐	☐
pas du tout acceptable	peu acceptable	assez acceptable	très acceptable	parfaitement acceptable

Mettez une croix sur la valeur choisie, en vous fiant à votre intuition. Essayez de varier vos réponses et d'utiliser toutes les possibilités de l'échelle.

Ne passez pas trop de temps sur chaque question, et ne revenez pas en arrière dans le questionnaire pour changer vos réponses. Par contre nous vous demandons de faire attention à bien mettre une note pour chaque continuation. Merci de votre collaboration.

Exemples

1 Pour un bon fonctionnement de l'entreprise, il y a lieu également *example1*

– d'accorder l'intérêt le plus soutenu possible aux élements qui améliorent la formation.

1 ☐ ☐ ☒ ☐ ☐ **5**

– d'accorder aux élements qui améliorent la formation l'intérêt le plus soutenu possible.

1 ☐ ☒ ☐ ☐ ☐ **5**

2 Les mesures pour l'emploi annoncées sur France 2, reprennent et amplifient des programmes existants ou en cours d'adaptation, *example2*

– avec le souci d'apporter le traitement le plus adapté à l'ensemble des situations.

1 ☐ ☐ ☒ ☐ ☐ **5**

– avec le souci d'apporter à l'ensemble des situations le traitement le plus adapté.

1 ☐ ☐ ☐ ☐ ☒ **5**

Test

1 Les affaires judiciaires de Marie ne s'arrangent pas. *JMAGT14*

– La plainte qu'avait exprès déposée son avocat sans consulter les délais de prescription fut classée sans suite.

1 ☐ ☐ ☐ ☐ ☐ **5**

– La plainte que son avocat avait exprès déposée sans consulter les délais de prescription fut classée sans suite.

1 ☐ ☐ ☐ ☐ ☐ **5**

2 Les affaires judiciaires de Marie ne s'arrangent pas. *JMAGT13*

– La plainte que son avocat avait bêtement déposée sans consulter les délais de prescription fut classée sans suite.

1 ☐ ☐ ☐ ☐ ☐ **5**

– La plainte qu'avait bêtement déposée son avocat sans consulter les délais de prescription fut classée sans suite.

1 ☐ ☐ ☐ ☐ ☐ **5**

3 Deux sondages fiables seulement ont été publiés depuis le débat de vendredi dernier. Le premier, mené dans la foulée de la rencontre télévisée, *AABCJT6*

– donne l'avantage à Kerry.

1 ☐ ☐ ☐ ☐ ☐ **5**

– donne à Kerry l'avantage.

1 ☐ ☐ ☐ ☐ ☐ **5**

4 La direction n'est pas correcte avec les jeunes recrutés. *JMAGT11*

– Les dix jours de congés que le nouvel employé a naïvement pris ont été interprétés comme une marque de paresse.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Les dix jours de congés qu'a naïvement pris le nouvel employé ont été interprétés comme une marque de paresse.

1 ☐ ☐ ☐ ☐ ☐ **5**

5 Guillaume en était déjà à sa troisième truite, c'est la première fois qu'il en prenait autant en si peu de temps. Fier de sa pêche, il se promenait sur la rive *AABCJT11*

– et montrait son précieux butin à son copain.

1 ☐ ☐ ☐ ☐ ☐ **5**

– et montrait à son copain son précieux butin.

1 ☐ ☐ ☐ ☐ ☐ **5**

6 Je dois amener ma voiture au contrôle anti-pollution, *JMAGT1*

– le moteur, qu'a gentiment révisé mon copain Pierre, devrait passer le test sans problème.

1 ☐ ☐ ☐ ☐ ☐ **5**

– le moteur, que mon copain Pierre a gentiment révisé, devrait passer le test sans problème.

1 ☐ ☐ ☐ ☐ ☐ **5**

7 Le déménagement s'est vraiment mal passé, *JMAGT4*

– le vase, qu'a inopinément renversé Pierre en déplaçant la commode, ne pourra pas être remplacé.

1 ☐ ☐ ☐ ☐ ☐ **5**

– le vase, que Pierre a inopinément renversé en déplaçant la commode, ne pourra pas être remplacé.

1 ☐ ☐ ☐ ☐ ☐ **5**

8 Il faut que les Israéliens maintenant, dans les prochaines semaines, dans les prochains mois, *AABCJT8*

– donnent à ces questions les réponses appropriées.

1 ☐ ☐ ☐ ☐ ☐ **5**

– donnent les réponses appropriées à ces questions.

1 ☐ ☐ ☐ ☐ ☐ **5**

9 La justice new-yorkaise a refusé *AABCJT1*

– d'accorder sa remise en liberté conditionnelle au meurtrier de John Lennon.

1 ☐ ☐ ☐ ☐ ☐ **5**

– d'accorder au meurtrier de John Lennon sa remise en liberté conditionnelle.

1 ☐ ☐ ☐ ☐ ☐ **5**

10 La vie du peintre Bartholomeo Renzi a été exceptionnellement facile. *JMAGT15*

– La pension que son mécène lui a généreusement versée lui a permis de poursuivre son œuvre sans soucis matériels.

1 ☐ ☐ ☐ ☐ ☐ **5**

– La pension que lui a généreusement versée son mécène lui a permis de poursuivre son œuvre sans soucis matériels.

1 ☐ ☐ ☐ ☐ ☐ **5**

11 Le déménagement s'est mal passé, *JMAGT3*

– le vase, que Pierre a maladroitement renversé en déplaçant la commode, ne pourra pas être remplacé.

1 ☐ ☐ ☐ ☐ ☐ **5**

– le vase, qu'a maladroitement renversé Pierre en déplaçant la commode, ne pourra pas être remplacé.

1 ☐ ☐ ☐ ☐ ☐ **5**

12 L'avenir du personnel de l'entreprise n'est pas désespéré. *JMAGT8*

– Le refus que l'intersyndicale a courageusement opposé à toute proposition de compensation financière laisse ouverte la perspective d'un procès honnête aux prud'hommes.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Le refus qu'a courageusement opposé l'intersyndicale à toute proposition de compensation financière laisse ouverte la perspective d'un procès honnête aux prud'hommes.

1 ☐ ☐ ☐ ☐ ☐ **5**

13 La vie du peintre Bartholomeo Renzi a été exceptionnellement facile. *JMAGT16*

– La pension que son mécène lui a régulièrement versée lui a permis de poursuivre son œuvre sans soucis matériels.

1 ☐ ☐ ☐ ☐ ☐ **5**

– La pension que lui a régulièrement versée son mécène lui a permis de poursuivre son œuvre sans soucis matériels.

1 ☐ ☐ ☐ ☐ ☐ **5**

14 Pierre est très ennuyé. *JMAGT10*

– Il a demandé à la police qu'on retrouve la valise que son fils avait à nouveau oubliée à la gare.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Il a demandé à la police qu'on retrouve la valise qu'avait à nouveau oubliée son fils à la gare.

1 ☐ ☐ ☐ ☐ ☐ **5**

15 A la pause, le score était toujours vierge *AABCJT5*

– et l'équipe de Charmes devait à son courageux gardien le maintien du score.

1 ☐ ☐ ☐ ☐ ☐ **5**

– et l'équipe de Charmes devait le maintien du score à son courageux gardien.

1 ☐ ☐ ☐ ☐ ☐ **5**

16 Je suis commis d'office en ce moment au tribunal des prud'hommes et j'en bave. *JMAGT5*

– Hier, j'ai accepté de défendre une secrétaire que les délégués syndicaux de son entreprise avaient lâchement abandonnée.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Hier, j'ai accepté de défendre une secrétaire qu'avaient lâchement abandonnée les délégués syndicaux de son entreprise.

1 ☐ ☐ ☐ ☐ ☐ **5**

17 Son procès fut une véritable épreuve. *JMAGT19*

– Il n'est pas prêt d'oublier la chanson que la salle a sadiquement entonnée quand il fut condamné à dix ans de prison.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Il n'est pas prêt d'oublier la chanson qu'a sadiquement entonnée la salle quand il fut condamné à dix ans de prison.

1 ☐ ☐ ☐ ☐ ☐ **5**

18 Je n'ai jamais connu ma famille paternelle. *JMAGT17*

– Mon père ne pardonna jamais le refus de solidarité financière que lui avaient égoïstement signifié ses frères lors de la faillite de son entreprise.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Mon père ne pardonna jamais le refus de solidarité financière que ses frères lui avaient égoïstement signifié lors de la faillite de son entreprise.

1 ☐ ☐ ☐ ☐ ☐ **5**

19 Recréer un véritable lien social passe par le souci *AABCJT12*

– de montrer son comportement exemplaire à l' extérieur.

1 ☐ ☐ ☐ ☐ ☐ **5**

– de montrer à l' extérieur son comportement exemplaire .

1 ☐ ☐ ☐ ☐ ☐ **5**

20 Des sources palestiniennes ont indiqué que l'armée a démoli une maison de trois étages, située près de la frontière avec l'Égypte à l'aide d'un bulldozer *AABCJT7*

– et sans donner à ses habitants le moindre préavis.

1 ☐ ☐ ☐ ☐ ☐ **5**

– et sans donner le moindre préavis à ses habitants.

1 ☐ ☐ ☐ ☐ ☐ **5**

21 Le groupe Bayard-Presse a bien résisté à la conjoncture déprimée de 2010. L'éditeur de La Croix - le quotidien, doté d'une nouvelle formule – *AABCJT3*

– doit cette relative solidité à sa diversification.

1 ☐ ☐ ☐ ☐ ☐ **5**

– doit à sa diversification cette relative solidité.

1 ☐ ☐ ☐ ☐ ☐ **5**

22 Son déplacement en province s'est finalement bien passé. *JMAGT21*

– La voiture de location qu'avait sagement réservée sa secrétaire l'attendait devant la gare complètement vide à la suite d'un appel à la grève générale.

1 ☐ ☐ ☐ ☐ ☐ **5**

– La voiture de location que sa secrétaire avait sagement réservée l'attendait devant la gare complètement vide à la suite d'un appel à la grève générale.

1 ☐ ☐ ☐ ☐ ☐ **5**

23 Je n'ai jamais connu ma famille paternelle. *JMAGT18*

– Mon père ne pardonna jamais le refus de solidarité financière que lui avaient sèchement signifié ses frères lors de la faillite de son entreprise.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Mon père ne pardonna jamais le refus de solidarité financière que ses frères lui avaient sèchement signifié lors de la faillite de son entreprise.

1 ☐ ☐ ☐ ☐ ☐ **5**

24 Les concurrents se penchent sur la structure du capital de Perrier *AABCJT16*

– dans l'espoir de prendre aux Agnelli l'affaire.

1 ☐ ☐ ☐ ☐ ☐ **5**

– dans l'espoir de prendre l'affaire aux Agnelli.

1 ☐ ☐ ☐ ☐ ☐ **5**

25 Son déplacement en province s'est finalement bien passé. *JMAGT22*

– La voiture de location qu'avait discrètement réservée sa secrétaire l'attendait devant la gare complètement vide à la suite d'un appel à la grève générale.

1 ☐ ☐ ☐ ☐ ☐ **5**

– La voiture de location que sa secrétaire avait discrètement réservée l'attendait devant la gare complètement vide à la suite d'un appel à la grève générale.

1 ☐ ☐ ☐ ☐ ☐ **5**

26 Condamné à une mort lente, Marcel *AABCJT4*

– doit sa libération inattendue aux troupes soviétiques.

1 ☐ ☐ ☐ ☐ ☐ **5**

– doit aux troupes soviétiques sa libération inattendue.

1 ☐ ☐ ☐ ☐ ☐ **5**

27 Pierre est très ennuyé. *JMAGT9*

– Il a demandé à la police qu'on retrouve la valise qu'avait stupidement oubliée son fils à la gare.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Il a demandé à la police qu'on retrouve la valise que son fils avait stupidement oubliée à la gare.

1 ☐ ☐ ☐ ☐ ☐ **5**

28 Je suis commis d'office en ce moment au tribunal des prud'hommes et j'en bave. *JMAGT6*

– Hier, j'ai accepté de défendre une secrétaire qu'avaient complètement abandonnée les délégués syndicaux de son entreprise.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Hier, j'ai accepté de défendre une secrétaire que les délégués syndicaux de son entreprise avaient complètement abandonnée.

1 ☐ ☐ ☐ ☐ ☐ **5**

29 Son procès fut une véritable épreuve. *JMAGT20*

– Il n'est pas prêt d'oublier la chanson qu'a bruyamment entonnée la salle quand il fut condamné à dix ans de prison.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Il n'est pas prêt d'oublier la chanson que la salle a bruyamment entonnée quand il fut condamné à dix ans de prison.

1 ☐ ☐ ☐ ☐ ☐ **5**

30 Le gel du programme du FMI risque *AABCJT13*

– de porter le coup de trop à notre politique économique.

1 ☐ ☐ ☐ ☐ ☐ **5**

– de porter à notre politique économique le coup de trop.

1 ☐ ☐ ☐ ☐ ☐ **5**

31 Je dois amener ma voiture au contrôle anti-pollution , *JMAGT2*

– le moteur, que mon copain Pierre a soigneusement révisé, devrait passer le test sans problème.

1 ☐ ☐ ☐ ☐ ☐ **5**

– le moteur, qu'a soigneusement révisé mon copain Pierre, devrait passer le test sans problème.

1 ☐ ☐ ☐ ☐ ☐ **5**

32 Une nuit d'été la mère de Pierre lui annonce qu'elle a arrêté la date de son mariage avec Lucie. Pierre fonce dans la nuit *AABCJT14*

– porter la bonne nouvelle à sa fiancée.

1 ☐ ☐ ☐ ☐ ☐ **5**

– porter à sa fiancée la bonne nouvelle.

1 ☐ ☐ ☐ ☐ ☐ **5**

33 L'avenir du personnel de l'entreprise n'est pas désespéré. *JMAGT7*

– Le refus qu'a intelligemment opposé l'intersyndicale à toute proposition de compensation financière laisse ouverte la perspective d'un procès honnête aux prud'hommes.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Le refus que l'intersyndicale a intelligemment opposé à toute proposition de compensation financière laisse ouverte la perspective d'un procès honnête aux prud'hommes.

1 ☐ ☐ ☐ ☐ ☐ **5**

34 Ce bon résultat concerne les marques Audi et Seat *AABCJT2*

– et doit à l'explosion des ventes en Allemagne l'essentiel de sa progression globale nette.

1 ☐ ☐ ☐ ☐ ☐ **5**

– et doit l'essentiel de sa progression globale nette à l'explosion des ventes en Allemagne.

1 ☐ ☐ ☐ ☐ ☐ **5**

35 Toutefois l'année dernière et celle d'avant, *AABCJT15*

– il était parvenu à prendre ce set à Agassi.

1 ☐ ☐ ☐ ☐ ☐ **5**

– il était parvenu à prendre à Agassi ce set.

1 ☐ ☐ ☐ ☐ ☐ **5**

36 J'essaie que mon personnage soit cohérent à travers les chansons et il y a toujours un lien avec moi *AABCJT10*

– qui donne aux chansons sa légitimité.

1 ☐ ☐ ☐ ☐ ☐ **5**

– qui donne sa légitimité aux chansons.

1 ☐ ☐ ☐ ☐ ☐ **5**

37 Ici, la terre et la mer se mélangent à l'infini. Seuls les flamants roses *AABCJT9*

– donnent à ce paysage le relief attendu.

1 ☐ ☐ ☐ ☐ ☐ **5**

– donnent le relief attendu à ce paysage.

1 ☐ ☐ ☐ ☐ ☐ **5**

38 La direction n'est pas correcte avec les jeunes recrutés. *JMAGT12*

– Les dix jours de congés que le nouvel employé a récemment pris furent interprétés comme une marque de paresse.

1 ☐ ☐ ☐ ☐ ☐ **5**

– Les dix jours de congés qu'a récemment pris le nouvel employé ont été interprétés comme une marque de paresse.

1 ☐ ☐ ☐ ☐ ☐ **5**

Commentaires libres

Quelles remarques avez-vous sur ce questionnaire ?

Avez-vous trouvé cela difficile ?

Avez-vous utilisé une stratégie ou une méthode pour répondre ?

Avez-vous remarqué quelque chose de particulier ?

Que pensez vous que soit le but de cette étude ?

Si vous voulez avoir les résultats, envoyez un mail à `abeille@linguist.jussieu.fr`

Annexe D

Questionnaire sur le statut *donné* ou *nouveau* du SP

Expérience de linguistique

Vous allez lire une ou plusieurs phrases suivies de deux continuations. Par exemple :

1 Les mesures pour l'emploi annoncées sur France 2, reprennent et amplifient des programmes existants ou en cours d'adaptation,

– avec le souci d'apporter le traitement le plus adapté à l'ensemble des situations.
– avec le souci d'apporter à l'ensemble des situations le traitement le plus adapté.

L'expérience porte sur les deux continuations proposées. Certaines continuations sont naturelles et appropriées dans un contexte normal d'usage de la langue. D'autres sont moins naturelles, parfois complètement inacceptables. En vous fiant à votre intuition, vous devez juger si ces continuations sont appropriées dans le contexte où elles apparaissent. Vous allez juger l'acceptabilité de ces phrases ou parties de phrase au moyen d'une échelle allant de 1 à 10. Vous noterez "1" une réponse qui n'est pas acceptable et "10" une phrase tout à fait acceptable. Par exemple :

2 Pour un bon fonctionnement de l'entreprise, il y a lieu également

– d'accorder l'intérêt le plus soutenu possible aux élements qui améliorent la formation.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
☐	☐	☐	☐	☐	☐	☐	☐	☐	☒
pas du tout acceptable		peu acceptable		plutôt pas acceptable	plutôt acceptable		très acceptable		parfaitement acceptable

– d'accorder aux élements qui améliorent la formation l'intérêt le plus soutenu possible.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
☐	☐	☐	⊛	☐	☐	☐	☐	☐	☐
pas du tout acceptable		peu acceptable		plutôt pas acceptable	plutôt acceptable		très acceptable		parfaitement acceptable

Si vous pensez que "d'accorder l'intérêt le plus soutenu possible aux élements qui améliorent la formation" est une partie de phrase parfaitement acceptable, vous cochez 10. Si vous pensez que "d'accorder aux élements qui améliorent la formation l'intérêt le plus soutenu possible" est une partie de phrase pas très acceptable, pas très naturelle dans ce contexte, vous cochez 4.

N'hésitez pas à utiliser la totalité de l'échelle dans vos jugements : vos jugements peuvent se situer aux extrêmes aussi bien qu'entre les extrêmes.

- Si vous le souhaitez, vous pouvez lire les phrases à voix basse avant de porter votre jugement.
- Certaines phrases pourront vous paraître "littéraires" ou d'un registre soutenu. En effet, on peut les trouver dans des journaux comme *Le Monde* ou dans des romans classiques. Ne tenez pas compte de cet aspect stylistique. Ne faites confiance qu'à votre réaction première.
- L'expérience ne porte pas sur la ponctuation. N'en tenez pas compte dans vos jugements.
- Ne passez pas trop de temps sur chaque question, et ne revenez pas en arrière dans le questionnaire pour changer vos réponses.
- Nous vous demandons de faire attention à bien mettre une note pour chaque continuation.

Avant de commencer, pouvez-vous répondre à quelques questions vous concernant ?

Quel est votre sexe ? F ☐ M ☐

Quel est votre âge ?

Dans quel(s) département(s) ou pays avez-vous passé votre enfance (jusqu'à 10 ans) ?

Quelle langue parliez-vous dans la famille avec vos parents avant d'être scolarisé ?

Parlez-vous couramment une ou plusieurs langues autres que le français ? Oui ☐ Non ☐

Si vous parlez plusieurs langues autres que le français :

- Laquelle parlez-vous le mieux ?
- L'avez-vous apprise dans la famille ? Oui ☐ Non ☐
- La pratiquez-vous de manière quotidienne en ce moment :
  - en la parlant ? Oui ☐ Non ☐
  - en l'écrivant ? Oui ☐ Non ☐
  - en la lisant ? Oui ☐ Non ☐

Merci !

Attention le questionnaire est imprimé recto-verso.

1 En Camargue, seuls les flamants roses peuvent *JTBCAA7*

– donner à un paysage monotone une pointe de relief.

1 2 3 4 5 6 7 8 9 10

– donner une pointe de relief à un paysage monotone.

1 2 3 4 5 6 7 8 9 10

2 L'entreprise demande au personnel d'encadrement de réduire son temps de travail tout en préservant la même charge de travail ! Forcément, il y a des erreurs, comme cette employée accusée *JTBCAA9*

– d'avoir laissé des produits périmés dans des rayons.

1 2 3 4 5 6 7 8 9 10

– d'avoir laissé dans des rayons des produits périmés.

1 2 3 4 5 6 7 8 9 10

3 Il a pris une chaise tout à fait banale et il nous a montré comment on peut *JTBCAA16*

– faire de cette simple chaise un objet d'art.

1 2 3 4 5 6 7 8 9 10

– faire un objet d'art de cette simple chaise.

1 2 3 4 5 6 7 8 9 10

4 Quand il était jeune, mon fils *JTBCAA6*

– portait à deux plantes exotiques un intérêt tout particulier.

1 2 3 4 5 6 7 8 9 10

– portait un intérêt tout particulier à deux plantes exotiques.

1 2 3 4 5 6 7 8 9 10

5 Un service de location de voitures sur le modèle de Vélib me serait très utile. *JMINV-ELAB7*

– N'a de voiture particulière aucun de mes voisins ou de mes amis les plus proches.

1 2 3 4 5 6 7 8 9 10

– N'ont de voiture particulière aucun de mes voisins ou de mes amis les plus proches.

1 2 3 4 5 6 7 8 9 10

6 Les affaires judiciaires de Marie ne s'arrangent pas. *JMAGT7*

– La plainte qu'avait exprès déposée son avocat sans consulter les délais de prescription fut classée sans suite.

1 2 3 4 5 6 7 8 9 10

– La plainte que son avocat avait exprès déposée sans consulter les délais de prescription fut classée sans suite.

1 2 3 4 5 6 7 8 9 10

7 Je suis sûr que, lui, c'est vraiment la personne idéale pour réussir à *JTBCAA15*

– faire d'une simple comédie romantique un très grand succès commercial.

1 2 3 4 5 6 7 8 9 10

– faire un très grand succès commercial d'une simple comédie romantique.

1 2 3 4 5 6 7 8 9 10

8 Les problèmes de sécurité se sont multipliés avec les pannes d'électricité à répétition. *JMINV-ELAB1*

– À vingt heures, ne pouvaient plus fonctionner normalement les ascenseurs et les caméras de surveillance.

1 2 3 4 5 6 7 8 9 10

– À vingt heures, ne pouvaient plus fonctionner normalement ni les ascenseurs ni les caméras de surveillance.

1 2 3 4 5 6 7 8 9 10

9 En plus de mes études, *JTBCAA13*

– je consacre de nombreuses heures à une association.

1 2 3 4 5 6 7 8 9 10

– je consacre à une association de nombreuses heures.

1 2 3 4 5 6 7 8 9 10

10 L' histoire de Monsieur Demaret mérite d'être racontée. Il y a deux ans, notre homme *JTBCAA11*

– trouve sur un trottoir de Nancy un billet de 100 euros.

1 2 3 4 5 6 7 8 9 10

– trouve un billet de 100 euros sur un trottoir de Nancy.

1 2 3 4 5 6 7 8 9 10

11 Depuis 1998, le taux d'imposition des terrains agricoles est de 4%. Hier, le conseil a décidé à l'unanimité *JTBCAA3*

– d'appliquer au taux de 1998 un coefficient de 1,5.

1 2 3 4 5 6 7 8 9 10

– d'appliquer un coefficient de 1,5 au taux de 1998.

1 2 3 4 5 6 7 8 9 10

12 Le week-end "football" a été riche en rebondissements. *JMINV-ELAB6*

– Ne se sont qualifiés, hier soir, pour la phase finale du championnat ni les nordistes de Lens ni les parisiens du PSG.

1 2 3 4 5 6 7 8 9 10

– Ne se sont pas qualifiés, hier soir, pour la phase finale du championnat les nordistes de Lens et les parisiens du PSG.

1 2 3 4 5 6 7 8 9 10

13 L'école primaire Vitruve, à Paris, a un fonctionnement particulier. Les instituteurs *JTBCAA4*

– appliquent un même principe pédagogique à chaque nouvel apprentissage.

1 2 3 4 5 6 7 8 9 10

– appliquent à chaque nouvel apprentissage un même principe pédagogique.

1 2 3 4 5 6 7 8 9 10

14 La fermeture de l'usine aura les répercussions qu'on connaît malheureusement trop bien. *JMINV-ELAB3*

– Dans l'année à venir, ne retrouveront pas d'emploi stable les seniors et les jeunes sans qualification.

1 2 3 4 5 6 7 8 9 10

– Dans l'année à venir, ne retrouveront un emploi stable ni les seniors ni les jeunes sans qualification.

1 2 3 4 5 6 7 8 9 10

15 Je dois amener ma voiture au contrôle anti-pollution , *JMAGT1*

– le moteur, que mon copain Pierre a soigneusement révisé, devrait passer le test sans problème.

1 2 3 4 5 6 7 8 9 10
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

– le moteur, qu'a soigneusement révisé mon copain Pierre, devrait passer le test sans problème.

1 2 3 4 5 6 7 8 9 10
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

16 Je suis sûre qu' il y a des gens qui accèdent à ce serveur pour espionner les autres, alors moi, j'ai décidé d'arrêter *JTBCAA10*

– de laisser sur le serveur des messages personnels.

1 2 3 4 5 6 7 8 9 10
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

– de laisser des messages personnels sur le serveur.

1 2 3 4 5 6 7 8 9 10
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

17 La vie du peintre Bartholomeo Renzi a été exceptionnellement facile. *JMAGT8*

– La pension que son mécène lui a généreusement versée lui a permis de poursuivre son œuvre sans soucis matériels.

1 2 3 4 5 6 7 8 9 10
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

– La pension que lui a généreusement versée son mécène lui a permis de poursuivre son œuvre sans soucis matériels.

1 2 3 4 5 6 7 8 9 10
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

18 Les résultats définitifs viennent de nous parvenir. *JMINV-ELAB8*

– N'ont réussi en session de rattrapage aucun étudiant d'informatique ou de linguistique.

1 2 3 4 5 6 7 8 9 10
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

– N'a réussi en session de rattrapage aucun étudiant d'informatique ou de linguistique.

1 2 3 4 5 6 7 8 9 10
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

19 L'identification des agresseurs impliqués dans le hold-up qui a eu lieu ce matin sera difficile. *JMINV-ELAB5*

– N'ont pas pu voir le visage du braqueur le bijoutier et le client.

1 2 3 4 5 6 7 8 9 10
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

– N'ont pu voir le visage du braqueur ni le bijoutier ni le client.

1 2 3 4 5 6 7 8 9 10

20 L'avenir du personnel de l'entreprise n'est pas désespéré. *JMAGT4*

– Le refus que l'intersyndicale a intelligemment opposé à toute proposition de compensation financière laisse ouverte la perspective d'un procès honnête aux prud'hommes.

1 2 3 4 5 6 7 8 9 10

– Le refus qu'a intelligemment opposé l'intersyndicale à toute proposition de compensation financière laisse ouverte la perspective d'un procès honnête aux prud'hommes.

1 2 3 4 5 6 7 8 9 10

21 Le 23 janvier, les chinois vont célébrer un grand évènement : le début de l'année du dragon. France Info *JTBCAA14*

– consacrera une journée spéciale à cet évènement.

1 2 3 4 5 6 7 8 9 10

– consacrera à cet évènement une journée spéciale.

1 2 3 4 5 6 7 8 9 10

22 Les objets rustiques ou artistiques proposés sur le marché Saint-Mathieu permettent, à peu de frais, *JTBCAA1*

– d'apporter à une décoration sobre une touche d'originalité.

1 2 3 4 5 6 7 8 9 10

– d'apporter une touche d'originalité à une décoration sobre.

1 2 3 4 5 6 7 8 9 10

23 Le gouvernement portugais tente de remonter la pente en proposant une politique économique basée sur la relance. Cependant, le gel du programme du FMI risque de *JTBCAA5*

– porter à cette politique économique un coup très dur.

1 2 3 4 5 6 7 8 9 10

– porter un coup très dur à cette politique économique.

1 2 3 4 5 6 7 8 9 10

24 Je suis commis d'office en ce moment au tribunal des prud'hommes et j'en bave. *JMAGT3*

– Hier, j'ai accepté de défendre une secrétaire que les délégués syndicaux de son entreprise avaient complètement abandonnée.

1 2 3 4 5 6 7 8 9 10

– Hier, j'ai accepté de défendre une secrétaire qu'avaient complètement abandonnée les délégués syndicaux de son entreprise.

1 2 3 4 5 6 7 8 9 10

25 La direction n'est pas correcte avec les jeunes recrutés. *JMAGT6*

– Les dix jours de congés que le nouvel employé a naïvement pris ont été interprétés comme une marque de paresse.

1 2 3 4 5 6 7 8 9 10

– Les dix jours de congés qu'a naïvement pris le nouvel employé ont été interprétés comme une marque de paresse.

1 2 3 4 5 6 7 8 9 10

26 Le verdict du tribunal est scandaleusement indulgent pour la direction. *JMINV-ELAB2*

– Selon le délibéré, n'ont pas commis de fautes graves la DRH et le directeur financier.

1 2 3 4 5 6 7 8 9 10

– Selon le délibéré, n'ont commis de fautes graves ni la DRH ni le directeur financier.

1 2 3 4 5 6 7 8 9 10

27 Depuis le début de la campagne, les différentes catégories d'électeurs changent fréquemment d'opinion. *JMINV-ELAB4*

– Néanmoins on peut constater qu'à trois jours de l'échéance n'ont pas varié dans leur choix les professions libérales et les retraités.

1 2 3 4 5 6 7 8 9 10

– Néanmoins on peut constater qu'à trois jours de l'échéance, n'ont varié dans leur choix ni les professions libérales ni les retraités.

1 2 3 4 5 6 7 8 9 10

28 Il y a en bas de chez moi, un café pas très sympathique et une librairie que j'adore. Tous les jours de l'année, *JTBCAA12*

– on trouve dans la librairie un accueil chaleureux.

1 2 3 4 5 6 7 8 9 10

– on trouve un accueil chaleureux dans la librairie.

1 2 3 4 5 6 7 8 9 10

29 Le déménagement s'est mal passé. *JMAGT2*

– Le vase, que Pierre a maladroitement renversé en déplaçant la commode, ne pourra pas être remplacé.

1 2 3 4 5 6 7 8 9 10

– Le vase, qu'a maladroitement renversé Pierre en déplaçant la commode, ne pourra pas être remplacé.

1 2 3 4 5 6 7 8 9 10

30 Le naufrage du chalutier a été d'une extrême rapidité. *JMINV-ELAB9*

– N'a survécu au drame aucun membre d'équipage ou officier.

1 2 3 4 5 6 7 8 9 10

– N'ont survécu au drame aucun membre d'équipage ou officier.

1 2 3 4 5 6 7 8 9 10

31 Pierre est très ennuyé. *JMAGT5*

– Il a demandé à la police qu'on retrouve la valise qu'avait à nouveau oubliée son fils à la gare.

1 2 3 4 5 6 7 8 9 10

– Il a demandé à la police qu'on retrouve la valise que son fils avait à nouveau oubliée à la gare.

1 2 3 4 5 6 7 8 9 10

32 Dans les hôpitaux du centre-ville, c'est la panique. En raison des bombardements, les organisations humanitaires n'ont pas pu *JTBCAA2*

– apporter du matériel médical stérilisé aux hôpitaux en pénurie.

1 2 3 4 5 6 7 8 9 10

– apporter aux hôpitaux en pénurie du matériel médical stérilisé.

1 2 3 4 5 6 7 8 9 10

33 De nombreuses questions se posent à propos de la situation économique du pays. Il faut que les candidats maintenant *JTBCAA8*

– donnent à ces questions des réponses appropriées.

1 2 3 4 5 6 7 8 9 10

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

– donnent des réponses appropriées à ces questions.

1 2 3 4 5 6 7 8 9 10

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Commentaires libres

Quelles remarques avez-vous sur ce questionnaire ?

Avez-vous trouvé cela difficile ?

Avez-vous utilisé une stratégie ou une méthode pour répondre ?

Avez-vous remarqué quelque chose de particulier ?

Que pensez vous que soit le but de cette étude ?

Si vous voulez avoir les résultats, envoyez un mail à `jthuilier@linguist.jussieu.fr`

Annexe E

Intercepts aléatoires relatifs aux adjectifs

E.1. Modèle Lexicalisé

La liste présentée dans cette section contient les intercepts aléatoires associés à la variable `ladj` dans le Modèle Lexicalisé qui est reproduit dans la table E.1.

Les intercepts aléatoires sont accompagnés du nombre d'occurrences dans la table de données, ainsi que des bornes supérieures et inférieures de l'intervalle de confiance à 95%. Les intercepts sont organisés par ordre croissant.

```
Effets aléatoires :
  Groupes  Nom          Variance  Ecart-type
  lem_adj  (Intercept)  4.934     2.2213
Nombre d'obs. : 13933 ; groupes : lem_adj, 1750

Effets fixes :
              Estimation  Erreur-type  valeur z  Pr(>|z|)
  (Intercept) -2.9059     0.0893       -32.54    <2e-16   ***
```

Table E.1.: Paramètres du Modèle Lexicalisé

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
britannique	-3.60973873	-3.97342451	-3.246052955	89
total	-2.73597031	-3.04102514	-2.430915481	49
rapide	-2.41781139	-2.74069773	-2.094925039	35
exceptionnel	-2.23757552	-2.57215377	-1.902997284	29
complet	-2.14025968	-2.63118104	-1.649338311	17
considérable	-2.02671046	-2.53265975	-1.520761168	15
moyen	-1.92763209	-2.05250867	-1.802755515	67
habituel	-1.89604983	-2.42073891	-1.371360745	13
extraordinaire	-1.87434408	-2.23703647	-1.511651682	20
sensible	-1.82366654	-2.10296944	-1.544363641	26
étroit	-1.82252470	-2.35850651	-1.286542903	12
immédiat	-1.74219966	-2.29118736	-1.193211959	11
présent	-1.65369337	-2.21787979	-1.089506950	10
riche	-1.65369337	-2.21787979	-1.089506950	10
moderne	-1.55515940	-2.13742400	-0.972894793	9
précédent	-1.54034946	-1.64140459	-1.439294323	64
actuel	-1.47923610	-1.55161743	-1.406854776	87
classique	-1.44405282	-2.04829651	-0.839809138	8
possible	-1.44405282	-2.04829651	-0.839809138	8
rigoureux	-1.44405282	-2.04829651	-0.839809138	8
seul2	-1.44405282	-2.04829651	-0.839809138	8
majeur	-1.42667694	-1.60363885	-1.249715029	33
nécessaire	-1.36903466	-1.68190455	-1.056164762	17
constant	-1.35184794	-1.76993718	-0.933758690	12
modeste	-1.25906122	-1.68946298	-0.828659463	11
net	-1.24155163	-1.32812727	-1.154975984	63
unique	-1.19740251	-1.41651963	-0.978285397	23
réel	-1.17162127	-1.30389767	-1.039344865	39
aléatoire	-1.16772734	-1.83504894	-0.500405747	6
insuffisant	-1.16772734	-1.83504894	-0.500405747	6
substantiel	-1.16772734	-1.83504894	-0.500405747	6
utile	-1.16772734	-1.83504894	-0.500405747	6
divers	-1.14980652	-1.48368430	-0.815928741	14
délicat	-1.06362179	-1.40683973	-0.720403864	13
difficile	-1.04249638	-1.23887859	-0.846114175	24
douloureux	-0.98830995	-1.70408500	-0.272534901	5
dramatique	-0.98830995	-1.70408500	-0.272534901	5

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
exact	-0.98830995	-1.70408500	-0.272534901	5
inéluctable	-0.98830995	-1.70408500	-0.272534901	5
coûteux	-0.96872880	-1.32302736	-0.614430232	12
brutal	-0.83675911	-1.13222663	-0.541291579	14
traditionnel	-0.82012536	-0.95367910	-0.686571631	33
amer	-0.76331951	-1.54975230	0.023113286	4
esthétique	-0.76331951	-1.54975230	0.023113286	4
fidèle	-0.76331951	-1.54975230	0.023113286	4
inégal	-0.76331951	-1.54975230	0.023113286	4
irrésistible	-0.76331951	-1.54975230	0.023113286	4
rentable	-0.76331951	-1.54975230	0.023113286	4
inquiétant	-0.75063197	-1.26699290	-0.234271028	7
prestigieux	-0.75063197	-1.26699290	-0.234271028	7
chaud	-0.56387847	-1.12160775	-0.006149183	6
malheureux	-0.56387847	-1.12160775	-0.006149183	6
médiocre	-0.56387847	-1.12160775	-0.006149183	6
profond	-0.49256390	-0.73122792	-0.253899885	16
contraignant	-0.46312220	-1.36384398	0.437599593	3
égal	-0.46312220	-1.36384398	0.437599593	3
fou	-0.46312220	-1.36384398	0.437599593	3
franc	-0.46312220	-1.36384398	0.437599593	3
louable	-0.46312220	-1.36384398	0.437599593	3
périlleux	-0.46312220	-1.36384398	0.437599593	3
prudent	-0.46312220	-1.36384398	0.437599593	3
soudain	-0.46312220	-1.36384398	0.437599593	3
ambitieux	-0.45040440	-0.88344387	-0.017364930	8
lourd	-0.44439879	-0.65676647	-0.232031115	18
récent	-0.42135698	-0.52109422	-0.321619747	40
important	-0.33230819	-0.37031704	-0.294299348	106
banal	-0.33103207	-0.94964494	0.287580798	5
ferme	-0.33103207	-0.94964494	0.287580798	5
sain	-0.33103207	-0.94964494	0.287580798	5
sévère	-0.28784114	-0.54841437	-0.027267914	14
vigoureux	-0.26108639	-0.73249337	0.210320604	7
lent	-0.26108639	-0.73249337	0.210320604	7
violent	-0.26108639	-0.73249337	0.210320604	7
dur	-0.16696143	-0.44249368	0.108570818	13
solide	-0.15078175	-0.39285965	0.091296157	15
ancien2	-0.15078175	-0.39285965	0.091296157	15

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
libre	-0.12787968	-0.32268979	0.066930434	19
relatif	-0.02979598	-0.28661223	0.227020273	14
dangereux	-0.02863736	-0.37419996	0.316925241	10
proche	-0.02769505	-0.44543574	0.390045642	8
strict	-0.02769505	-0.44543574	0.390045642	8
remarquable	-0.02625517	-0.55428663	0.501776284	6
heureux	-0.02625517	-0.55428663	0.501776284	6
sombre	-0.02625517	-0.55428663	0.501776284	6
apparent	-0.02378226	-0.74123375	0.693669235	4
écrasant	-0.02378226	-0.74123375	0.693669235	4
flamboyant	-0.02378226	-0.74123375	0.693669235	4
impressionnant	-0.02378226	-0.74123375	0.693669235	4
inexorable	-0.02378226	-0.74123375	0.693669235	4
précieux	-0.02378226	-0.74123375	0.693669235	4
pur1	-0.02378226	-0.74123375	0.693669235	4
ample	-0.01854270	-1.13733904	1.100253646	2
brusque	-0.01854270	-1.13733904	1.100253646	2
confortable	-0.01854270	-1.13733904	1.100253646	2
dit	-0.01854270	-1.13733904	1.100253646	2
éminent	-0.01854270	-1.13733904	1.100253646	2
fin	-0.01854270	-1.13733904	1.100253646	2
flagrant	-0.01854270	-1.13733904	1.100253646	2
influent	-0.01854270	-1.13733904	1.100253646	2
interminable	-0.01854270	-1.13733904	1.100253646	2
irréversible	-0.01854270	-1.13733904	1.100253646	2
légendaire	-0.01854270	-1.13733904	1.100253646	2
lointain	-0.01854270	-1.13733904	1.100253646	2
luxueux	-0.01854270	-1.13733904	1.100253646	2
merveilleux	-0.01854270	-1.13733904	1.100253646	2
minuscule	-0.01854270	-1.13733904	1.100253646	2
mirobolant	-0.01854270	-1.13733904	1.100253646	2
mystérieux	-0.01854270	-1.13733904	1.100253646	2
pertinent	-0.01854270	-1.13733904	1.100253646	2
pesant	-0.01854270	-1.13733904	1.100253646	2
quelconque	-0.01854270	-1.13733904	1.100253646	2
regrettable	-0.01854270	-1.13733904	1.100253646	2
sage	-0.01854270	-1.13733904	1.100253646	2
salutaire	-0.01854270	-1.13733904	1.100253646	2
sournois	-0.01854270	-1.13733904	1.100253646	2

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
puissant	0.13036264	-0.18958156	0.450306848	11
indispensable	0.20710474	-0.26485338	0.679062856	7
énorme	0.26298091	0.06461593	0.461345887	19
bref	0.28090185	-0.33886007	0.900663757	5
différent	0.31353840	0.23521931	0.391857500	51
prochain	0.32712981	0.29371308	0.360546527	122
sérieux	0.32955169	0.20664727	0.452456109	32
dernier	0.41456963	0.40355493	0.425584333	380
large	0.41915394	0.23342645	0.604881421	21
intense	0.42038176	-0.48331807	1.324081578	3
probable	0.42038176	-0.48331807	1.324081578	3
fort	0.42735942	0.38926230	0.465456545	109
faible	0.63658444	0.54684018	0.726328709	48
étrange	0.71760782	-0.07278858	1.508004212	4
redoutable	0.71760782	-0.07278858	1.508004212	4
court	0.75731707	0.62359763	0.891036519	33
bas	0.79273917	0.69610001	0.889378326	47
grave	0.83322694	0.59196208	1.074491806	18
vif	0.89734918	0.72090578	1.073792568	26
extrême	0.94075807	0.22065801	1.660858139	5
inévitable	0.94075807	0.22065801	1.660858139	5
sacro-saint	0.94075807	0.22065801	1.660858139	5
gigantesque	0.94075807	0.22065801	1.660858139	5
rare	0.94499021	0.66497830	1.225002128	16
long	0.97521070	0.91295171	1.037469681	80
excellent	0.98525906	0.51976315	1.450754974	9
multiple	1.01858638	0.74503217	1.292140596	17
faux	1.11891198	0.44713825	1.790685717	6
vrai	1.17096008	0.84377468	1.498145475	15
futur	1.26634734	1.10726325	1.425431429	35
propre1	1.31072121	1.18479887	1.436643546	46
léger	1.32919719	1.15022518	1.508169198	32
éventuel	1.36695960	1.16020835	1.573710836	28
formidable	1.38044554	0.97057736	1.790313716	13
célèbre	1.39359059	0.78488567	2.002295499	8
juste	1.39359059	0.78488567	2.002295499	8
prétendu	1.39359059	0.78488567	2.002295499	8
plein	1.48794239	1.18826178	1.787622988	20
ultime	1.50412275	0.91743682	2.090808683	9

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
principal	1.52045553	1.46297372	1.577937342	118
jeune	1.76659472	1.39719938	2.135990058	19
nouveau	1.89811171	1.87877059	1.917452836	464
simple1	2.11111423	1.76962672	2.452601733	27
meilleur	2.17246308	1.99604338	2.348882794	59
certain	2.25619655	2.08199964	2.430393456	64
moindre	2.35997049	2.03509611	2.684844865	35
nombreux	2.44120697	2.19414647	2.688267472	51
haut	2.50075726	2.05289567	2.948618853	27
véritable	2.91229597	2.61675595	3.207835990	63
mauvais	2.93296789	2.52370518	3.342230606	44
petit	3.12712257	2.90256832	3.351676827	104
seul1	3.68995004	3.33142947	4.048470623	104
bon	3.81583683	3.46427975	4.167393903	120
autre	4.34591909	4.02038612	4.671452049	219
premier	4.60047282	4.36060841	4.840337224	396
grand	4.64156265	4.32864813	4.954477169	306

E.2. Modèle Global

Dans cette section sont présentés les intercepts aléatoires associés à la variable `ladj` dans le Modèle Global qui est reproduit dans la table E.2. Les intercepts aléatoires sont accompagnés du nombre d'occurrences dans la table de données, ainsi que des bornes supérieures et inférieures de l'intervalle de confiance à 95%. Les intercepts sont organisés par ordre croissant.

Effets aléatoires :

Groupes	Nom	Variance	Ecart-type
lem_adj	(Intercept)	2.1367	1.4618

Nombre d'obs. : 4994 ; groupes : lem_adj, 171

Effets fixes :

	Estimation	Erreur-type	valeur z	Pr(>|z|)	
(Intercept)	-0.29327	0.15559	-1.885	0.059452	
artDef=1	0.38677	0.11436	3.382	0.000720	***
detDem=1	1.53952	0.28554	5.392	6.98e-08	***
detPoss=1	0.96312	0.25320	3.804	0.000142	***
coord=1	-1.35389	0.28564	-4.740	2.14e-06	***
adjAnt=1	0.55510	0.26261	2.114	0.034531	*
adjPost=1	0.58506	0.16017	3.653	0.000259	***
sprep=1	0.86851	0.11127	7.805	5.94e-15	***
adv=1	-1.70414	0.18981	-8.978	< 2e-16	***
collocNA	-0.44146	0.02207	-20.006	< 2e-16	***
collocAN	0.36458	0.01997	18.256	< 2e-16	***

Corrélation des effets fixes :

	(Int)	def	dem	poss	coo	aAnt	aPos	sp	adv	NA
def=1	-.270									
dem=1	-.131	.209								
poss=1	-.149	.236	.093							
coord=1	-.099	.013	-.016	-.003						
aAnt=1	-.035	.003	.050	.013	.008					
aPost=1	-.196	-.037	.020	.028	.038	.016				
sprep=1	-.293	-.102	.052	.046	.007	.028	.173			
adv=1	-.064	-.075	-.009	-.069	.050	-.028	-.052	-.026		
colNA	-.119	-.050	-.028	-.021	.065	-.077	.072	.028	.126	
colAN	-.165	.083	.001	.034	-.057	.032	.097	.083	-.129	-.508

Table E.2.: Paramètres du Modèle Global

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
britannique	-3.4155593446	-3.759733851	-3.071384838	89
total	-2.4446065568	-2.790309644	-2.098903470	49
rapide	-2.0933962790	-2.424348696	-1.762443862	35
habituel	-2.0882467790	-2.581642817	-1.594850741	13
possible	-1.9521191882	-2.531436908	-1.372801469	8
majeur	-1.9476313725	-2.317229492	-1.578033253	33
extraordinaire	-1.9438324201	-2.335646599	-1.552018241	20
exceptionnel	-1.9229591935	-2.272124640	-1.573793747	29
moyen	-1.8500696401	-2.027842262	-1.672297018	67
complet	-1.7804033490	-2.310907050	-1.249899648	17
immédiat	-1.7766768600	-2.527543844	-1.025809876	11
présent	-1.7161928861	-2.281839941	-1.150545832	10
considérable	-1.7092450062	-2.252619674	-1.165870338	15
actuel	-1.5746884144	-1.665351030	-1.484025799	87
traditionnel	-1.5173145507	-1.667259018	-1.367370084	33
classique	-1.4974371486	-2.078660061	-0.916214236	8
regrettable	-1.4913874861	-2.871803552	-0.110971421	2
réel	-1.4689699776	-1.629837301	-1.308102654	39
moderne	-1.4522248644	-2.048369643	-0.856080085	9
net	-1.3632282250	-1.510202043	-1.216254407	63
irrésistible	-1.3511338008	-2.227115721	-0.475151881	4
unique	-1.3030669116	-1.579985440	-1.026148383	23
divers	-1.2797803488	-1.650759448	-0.908801249	14
sensible	-1.1989501016	-1.500963132	-0.896937071	26
amer	-1.1662387238	-2.011153053	-0.321324394	4
dramatique	-1.0904424966	-1.817601516	-0.363283477	5
étroit	-1.0745936186	-1.712074123	-0.437113114	12
exact	-1.0571984482	-1.908361104	-0.206035792	5
utile	-1.0148222313	-1.712589586	-0.317054877	6
rigoureux	-0.9597051799	-1.706781479	-0.212628881	8
violent	-0.9418975030	-1.510283051	-0.373511955	7
délicat	-0.9311726773	-1.393694594	-0.468650761	13
soudain	-0.9019944848	-1.839669137	0.035680167	3
brutal	-0.8750830337	-1.228064498	-0.522101569	14
substantiel	-0.8113967589	-1.531963883	-0.090829635	6
profond	-0.8086358306	-1.152484907	-0.464786754	16
nécessaire	-0.7665704228	-1.158005822	-0.375135024	17
inéluctable	-0.7635601776	-1.535412428	0.008292072	5
relatif	-0.7499917470	-1.109101467	-0.390882027	14
pertinent	-0.7481016890	-1.888285628	0.392082250	2

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
périlleux	-0.7412497108	-1.616467982	0.133968561	3
prestigieux	-0.7342514428	-1.382468179	-0.086034707	7
modeste	-0.6655181934	-1.174740599	-0.156295788	11
aléatoire	-0.6622707393	-1.364165927	0.039624448	6
chaud	-0.6550126092	-1.341397023	0.031371805	6
lourd	-0.6334966574	-1.038906504	-0.228086811	18
constant	-0.6220489054	-1.142143113	-0.101954698	12
proche	-0.6205229513	-1.282059346	0.041013444	8
récent	-0.6129387483	-0.746642189	-0.479235307	40
coûteux	-0.5962797025	-1.007149544	-0.185409862	12
difficile	-0.5796223314	-0.830443555	-0.328801108	24
louable	-0.5491762349	-1.603531859	0.505179389	3
précédent	-0.5008837512	-0.672966376	-0.328801126	64
ambitieux	-0.4879557872	-1.044343748	0.068432174	8
libre	-0.4596629762	-0.756116813	-0.163209139	19
luxueux	-0.4531008745	-1.511488325	0.605286576	2
inquiétant	-0.4321107286	-1.021322524	0.157101067	7
confortable	-0.3897373230	-1.584296377	0.804821731	2
insuffisant	-0.3806662678	-1.250988594	0.489656058	6
douloureux	-0.3712229241	-1.163332554	0.420886706	5
court	-0.3535214539	-0.590093653	-0.116949254	33
lointain	-0.3459441760	-1.530847889	0.838959537	2
inexorable	-0.3440515482	-1.041040691	0.352937594	4
vigoureux	-0.3201846108	-0.854262630	0.213893409	7
heureux	-0.2914289229	-0.885599980	0.302742134	6
ancien2	-0.2893731333	-0.599328858	0.020582592	15
malheureux	-0.2837680494	-0.906852180	0.339316081	6
riche	-0.2728047584	-1.204860929	0.659251412	10
mystérieux	-0.2657832207	-1.331710040	0.800143599	2
légendaire	-0.2657832207	-1.331710040	0.800143599	2
médiocre	-0.2404127719	-0.882854883	0.402029339	6
interminable	-0.2339546087	-1.273446030	0.805536813	2
quelconque	-0.2233925333	-1.264991576	0.818206509	2
esthétique	-0.2088365936	-1.250054071	0.832380884	4
impressionnant	-0.2006546883	-0.915436786	0.514127409	4
contraignant	-0.1880158704	-1.199982634	0.823950893	3
minuscule	-0.1644679451	-1.252061768	0.923125878	2
ferme	-0.1582205304	-0.886415784	0.569974723	5
fidèle	-0.1504749886	-1.001370244	0.700420267	4
prudent	-0.1408587944	-1.097451938	0.815734349	3
remarquable	-0.1231510418	-0.666730915	0.420428832	6

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
franc	-0.1026836636	-1.024688825	0.819321497	3
banal	-0.1015139345	-0.755387661	0.552359792	5
fort	-0.0971921781	-0.159946566	-0.034437790	109
long	-0.0859661455	-0.173865111	0.001932820	80
sage	-0.0793200188	-1.328075920	1.169435883	2
fin	-0.0711620594	-1.129754149	0.987430031	2
merveilleux	-0.0711620594	-1.129754149	0.987430031	2
irréversible	-0.0711620594	-1.129754149	0.987430031	2
sournois	-0.0711620594	-1.129754149	0.987430031	2
dit	-0.0485628226	-1.244870562	1.147744917	2
puissant	-0.0420103186	-0.475035117	0.391014480	11
seul2	-0.0019427787	-0.752896535	0.749010978	8
brusque	0.0003792981	-1.044053852	1.044812448	2
intense	0.0082560399	-0.943801766	0.960313846	3
mirobolant	0.0122912110	-1.146136040	1.170718462	2
probable	0.0242500874	-0.868441260	0.916941434	3
sévère	0.0361913596	-0.348120087	0.420502806	14
éminent	0.0403383205	-0.999044758	1.079721399	2
pesant	0.0513552944	-0.986986657	1.089697245	2
salutaire	0.1074901533	-1.009932159	1.224912466	2
bref	0.1333514814	-0.659055653	0.925758615	5
dur	0.1441313897	-0.202014306	0.490277085	13
grave	0.1682149403	-0.158216914	0.494646795	18
sain	0.1696658063	-0.608093192	0.947424805	5
rentable	0.1798462359	-0.692395276	1.052087748	4
écrasant	0.1885798268	-0.534862853	0.912022506	4
large	0.1929510019	-0.056400833	0.442302836	21
différent	0.2188087091	0.103723106	0.333894313	51
apparent	0.2201822231	-0.547686865	0.988051311	4
énorme	0.2376087344	0.013553892	0.461663577	19
flamboyant	0.2408574491	-0.596444315	1.078159214	4
lent	0.2470799076	-0.296128682	0.790288497	7
précieux	0.2514198972	-0.659667601	1.162507395	4
inégal	0.2786213925	-0.611575568	1.168818353	4
fou	0.2958494400	-0.783504822	1.375203702	3
important	0.3054749447	0.253804942	0.357144947	106
strict	0.3103896982	-0.285370441	0.906149838	8
égal	0.3114765177	-1.289952223	1.912905258	3
bas	0.3401602151	0.214619839	0.465700591	47
pur1	0.3485686379	-0.683856889	1.380994165	4
influent	0.3513767601	-1.118705654	1.821459175	2

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
faible	0.3536855112	0.240455501	0.466915522	48
sombre	0.3905116014	-0.400315442	1.181338645	6
dangereux	0.4167273918	-0.100676973	0.934131757	10
sérieux	0.4289382166	0.258764353	0.599112080	32
flagrant	0.4304072314	-0.685182987	1.545997450	2
gigantesque	0.5284301882	-0.399168150	1.456028526	5
indispensable	0.5524787975	0.055099667	1.049857928	7
inévitable	0.5825473072	-0.186974822	1.352069437	5
solide	0.6114311235	0.302909259	0.919952988	15
ultime	0.6522497969	-0.009756921	1.314256515	9
futur	0.6799035399	0.508413966	0.851393113	35
célèbre	0.6865396815	-0.075965082	1.449044445	8
sacro-saint	0.7099165466	-0.003847464	1.423680557	5
ample	0.7326443636	-0.371583582	1.836872309	2
propre1	0.7418189504	0.561144565	0.922493336	46
formidable	0.7636426425	0.307966902	1.219318383	13
multiple	0.7645856320	0.455233267	1.073937997	17
léger	0.7911929259	0.539273309	1.043112543	32
vif	0.8092633282	0.579230523	1.039296133	26
prétendu	0.8134028782	0.156756148	1.470049608	8
vrai	0.8274687443	0.479132696	1.175804792	15
principal	0.8497670015	0.783165178	0.916368825	118
éventuel	0.8900675425	0.660328763	1.119806322	28
faux	0.9560753772	0.172012461	1.740138293	6
excellent	0.9998720051	0.417759608	1.581984403	9
extrême	1.0297065952	0.316848305	1.742564885	5
étrange	1.0510345279	-0.042013165	2.144082220	4
juste	1.1083641418	0.056438501	2.160289783	8
plein	1.2662552743	0.670406166	1.862104382	20
redoutable	1.3456418421	0.403891558	2.287392126	4
nouveau	1.4474534912	1.424716628	1.470190354	464
dernier	1.5025880987	1.474637819	1.530538379	380
rare	1.6706917280	1.152403707	2.188979749	16
jeune	1.6794532188	1.099216491	2.259689947	19
meilleur	1.7041681498	1.438771507	1.969564793	59
certain	1.9010752634	1.684861082	2.117289445	64
mauvais	2.0733644283	1.610057231	2.536671626	44
prochain	2.0839961243	2.017524488	2.150467761	122
haut	2.1453959490	1.670182850	2.620609048	27
véritable	2.1757525663	1.870710896	2.480794237	63
simple1	2.1968426521	1.818367878	2.575317427	27
nombreux	2.2120558990	1.956507162	2.467604636	51

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
petit	2.3028939607	2.063468646	2.542319275	104
bon	2.8327936709	2.443920915	3.221666427	120
seul1	2.9014019597	2.534442848	3.268361072	104
moindre	3.3834627622	2.900127744	3.866797781	35
premier	3.4393436219	3.185806319	3.692880925	396
autre	3.7623594178	3.417037922	4.107680913	219
grand	4.0476945588	3.739919056	4.355470061	306

E.3. Données relatives aux 171 adjectifs alternant

E.3.1. Nombre d'occurrences antéposées et postposées par corpus

E.3.1.1. FTB

Adjectif	total	ant	post
actuel	87	16	71
aléatoire	6	1	5
ambitieux	8	3	5
amer	4	1	3
ample	2	1	1
ancien	15	7	8
apparent	4	2	2
autre	219	218	1
banal	5	2	3
bas	47	33	14
bon	120	119	1
bref	5	3	2
britannique	89	1	88
brusque	2	1	1
brutal	14	4	10
célèbre	8	7	1
certain	64	59	5
chaud	6	2	4
classique	8	1	7
complet	17	1	16
confortable	2	1	1
considérable	15	1	14
constant	12	2	10
contraignant	3	1	2
court	33	23	10
coûteux	12	3	9
dangereux	10	5	5
délicat	13	3	10
dernier	380	232	148
différent	51	30	21
difficile	24	6	18
dit	2	1	1
divers	14	3	11
douloureux	5	1	4
dramatique	5	1	4
dur	13	6	7

Adjectif	total	ant	post
écrasant	4	2	2
égal	3	1	2
éminent	2	1	1
énorme	19	11	8
esthétique	4	1	3
étrange	4	3	1
étroit	12	1	11
éventuel	28	23	5
exact	5	1	4
excellent	9	7	2
exceptionnel	29	2	27
extraordinaire	20	2	18
extrême	5	4	1
faible	48	32	16
faux	6	5	1
ferme	5	2	3
fidèle	4	1	3
fin	2	1	1
flagrant	2	1	1
flamboyant	4	2	2
formidable	13	11	2
fort	109	67	42
fou	3	1	2
franc	3	1	2
futur	35	28	7
gigantesque	5	4	1
grand	306	305	1
grave	18	13	5
habituel	13	1	12
haut	27	26	1
heureux	6	3	3
immédiat	11	1	10
important	106	45	61
impressionnant	4	2	2
indispensable	7	4	3
inégal	4	1	3

Adjectif	total	ant	post
inéluctable	5	1	4
inévitable	5	4	1
inexorable	4	2	2
influent	2	1	1
inquiétant	7	2	5
insuffisant	6	1	5
intense	3	2	1
interminable	2	1	1
irrésistible	4	1	3
irréversible	2	1	1
jeune	19	17	2
juste	8	7	1
large	21	13	8
légendaire	2	1	1
léger	32	26	6
lent	7	3	4
libre	19	9	10
lointain	2	1	1
long	80	59	21
louable	3	1	2
lourd	18	7	11
luxueux	2	1	1
majeur	33	6	27
malheureux	6	2	4
mauvais	44	43	1
médiocre	6	2	4
meilleur	59	54	5
merveilleux	2	1	1
minuscule	2	1	1
mirobolant	2	1	1
moderne	9	1	8
modeste	11	2	9
moindre	35	33	2
moyen	67	8	59
multiple	17	13	4
mystérieux	2	1	1
nécessaire	17	3	14
net	63	14	49
nombreux	51	48	3
nouveau	464	406	58
périlleux	3	1	2
pertinent	2	1	1
pesant	2	1	1
petit	104	101	3
plein	20	17	3
possible	8	1	7
précédent	64	11	53
précieux	4	2	2
premier	396	394	2
présent	10	1	9
prestigieux	7	2	5
prétendu	8	7	1
principal	118	98	20
probable	3	2	1
prochain	122	72	50
proche	8	4	4
profond	16	6	10
propre	46	37	9
prudent	3	1	2
puissant	11	6	5
pur	4	2	2
quelconque	2	1	1
rapide	35	2	33
rare	16	12	4
récent	40	16	24
redoutable	4	3	1
réel	39	9	30
regrettable	2	1	1
relatif	14	7	7
remarquable	6	3	3
rentable	4	1	3
riche	10	1	9
rigoureux	8	1	7
sacro-saint	5	4	1
sage	2	1	1
sain	5	2	3
salutaire	2	1	1
sensible	26	3	23
sérieux	32	19	13
seul	112	104	8
sévère	14	6	8
simple	27	25	2
solide	15	7	8
sombre	6	3	3
soudain	3	1	2
sournois	2	1	1
strict	8	4	4
substantiel	6	1	5
total	49	2	47

Adjectif	total	ant	post
traditionnel	33	10	23
ultime	9	8	1
unique	23	5	18
utile	6	1	5
véritable	63	61	2
vif	26	19	7
vigoureux	7	3	4
violent	7	3	4
vrai	15	12	3

E.3.1.2. ESTER

Adjectif	total	ant	post
actuel	29	9	20
aléatoire	1	0	1
ambitieux	0	0	0
amer	3	0	3
ample	0	0	0
ancien	99	90	9
apparent	2	1	1
autre	100	99	1
banal	4	0	4
bas	9	7	2
bon	100	98	2
bref	1	1	0
britannique	71	0	71
brusque	1	1	0
brutal	4	0	4
célèbre	15	11	4
certain	53	53	0
chaud	9	1	8
classique	12	1	11
complet	13	1	12
confortable	1	1	0
considérable	9	1	8
constant	4	1	3
contraignant	0	0	0

Adjectif	total	ant	post
court	16	7	9
coûteux	1	1	0
dangereux	12	3	9
délicat	6	2	4
dernier	100	78	22
différent	31	14	17
difficile	19	1	18
dit	1	0	1
divers	7	1	6
douloureux	3	0	3
dramatique	0	0	0
dur	5	0	5
écrasant	1	0	1
égal	1	0	1
éminent	1	0	1
énorme	11	11	0
esthétique	0	0	0
étrange	3	2	1
étroit	5	1	4
éventuel	23	18	5
exact	6	0	6
excellent	10	8	2
exceptionnel	11	0	11
extraordinaire	9	1	8

Adjectif	total	ant	post
extrême	7	3	4
faible	4	2	2
faux	8	8	0
ferme	3	0	3
fidèle	3	2	1
fin	8	2	6
flagrant	0	0	0
flamboyant	0	0	0
formidable	1	1	0
fort	60	35	25
fou	2	1	1
franc	2	1	1
futur	36	35	1
gigantesque	3	2	1
grand	99	96	3
grave	19	5	14
habituel	9	0	9
haut	26	26	0
heureux	1	0	1
immédiat	11	0	11
important	71	11	60
impressionnant	3	0	3
indispensable	3	1	2
inégal	0	0	0
inéluctable	0	0	0
inévitable	5	1	4
inexorable	0	0	0
influent	2	0	2
inquiétant	7	2	5
insuffisant	0	0	0
intense	6	4	2
interminable	0	0	0
irrésistible	1	0	1
irréversible	0	0	0
jeune	58	56	2
juste	11	6	5
large	5	4	1
légendaire	0	0	0
léger	9	6	3
lent	3	0	3
libre	16	3	13
lointain	1	0	1
long	18	14	4

Adjectif	total	ant	post
louable	0	0	0
lourd	17	3	14
luxueux	2	1	1
majeur	22	2	20
malheureux	1	0	1
mauvais	27	27	0
médiocre	6	0	6
meilleur	37	36	1
merveilleux	1	0	1
minuscule	2	1	1
mirobolant	1	0	1
moderne	7	0	7
modeste	7	2	5
moindre	8	8	0
moyen	7	3	4
multiple	6	3	3
mystérieux	6	2	4
nécessaire	6	1	5
net	6	3	3
nombreux	43	40	3
nouveau	99	95	4
périlleux	1	0	1
pertinent	0	0	0
pesant	1	0	1
petit	99	99	0
plein	23	19	4
possible	20	5	15
précédent	17	3	14
précieux	2	0	2
premier	100	100	0
présent	4	0	4
prestigieux	3	3	0
prétendu	0	0	0
principal	83	63	20
probable	12	2	10
prochain	66	30	36
proche	3	1	2
profond	10	1	9
propre	40	36	4
prudent	1	0	1
puissant	4	2	2
pur	2	0	2
quelconque	1	1	0

Adjectif	total	ant	post
rapide	13	1	12
rare	11	8	3
récent	12	3	9
redoutable	3	1	2
réel	17	5	12
regrettable	1	0	1
relatif	2	0	2
remarquable	5	2	3
rentable	0	0	0
riche	9	3	6
rigoureux	2	1	1
sacro-saint	0	0	0
sage	0	0	0
sain	0	0	0
salutaire	0	0	0
sensible	12	0	12
sérieux	12	2	10
seul	77	77	0
sévère	5	0	5
simple	32	23	9
solide	4	0	4
sombre	5	0	5
soudain	0	0	0
sournois	0	0	0
strict	1	0	1
substantiel	0	0	0
total	26	3	23
traditionnel	9	2	7
ultime	3	3	0
unique	17	4	13
utile	6	0	6
véritable	48	48	0
vif	6	3	3
vigoureux	1	0	1
violent	15	5	10
vrai	43	43	0

E.3.1.3. CORAL-ROM

Adjectif	total	ant	post
actuel	10	1	9
aléatoire	0	0	0
ambitieux	1	1	0
amer	1	0	1
ample	0	0	0
ancien	19	15	4
apparent	2	1	1
autre	100	99	1
banal	4	0	4
bas	2	1	1
bon	45	45	0
bref	2	2	0
britannique	1	0	1
brusque	0	0	0
brutal	1	0	1
célèbre	4	2	2
certain	37	36	1
chaud	0	0	0
classique	6	0	6
complet	5	0	5
confortable	0	0	0
considérable	2	0	2
constant	1	0	1
contraignant	0	0	0
court	2	2	0
coûteux	1	0	1
dangereux	4	0	4
délicat	3	0	3
dernier	58	35	23
différent	39	13	26
difficile	8	0	8
dit	0	0	0
divers	2	1	1
douloureux	2	2	0
dramatique	3	0	3
dur	11	0	11
écrasant	2	0	2
égal	2	0	2
éminent	1	1	0
énorme	15	4	11
esthétique	1	0	1

Adjectif	total	ant	post
étrange	4	1	3
étroit	1	1	0
éventuel	1	0	1
exact	5	0	5
excellent	2	1	1
exceptionnel	3	0	3
extraordinaire	11	0	11
extrême	4	2	2
faible	3	1	2
faux	7	7	0
ferme	0	0	0
fidèle	1	0	1
fin	2	0	2
flagrant	1	0	1
flamboyant	1	0	1
formidable	2	2	0
fort	9	3	6
fou	2	0	2
franc	1	0	1
futur	3	2	1
gigantesque	1	0	1
grand	100	95	5
grave	5	0	5
habituel	3	0	3
haut	13	10	3
heureux	2	1	1
immédiat	2	0	2
important	44	6	38
impressionnant	2	0	2
indispensable	2	0	2
inégal	0	0	0
inéluctable	1	0	1
inévitable	0	0	0
inexorable	1	0	1
influent	0	0	0
inquiétant	0	0	0
insuffisant	0	0	0
intense	2	0	2
interminable	0	0	0
irrésistible	0	0	0
irréversible	0	0	0

Adjectif	total	ant	post
jeune	32	30	2
juste	1	1	0
large	5	0	5
légendaire	0	0	0
léger	8	7	1
lent	3	0	3
libre	12	6	6
lointain	0	0	0
long	19	13	6
louable	0	0	0
lourd	5	1	4
luxueux	0	0	0
majeur	9	1	8
malheureux	2	1	1
mauvais	19	19	0
médiocre	0	0	0
meilleur	22	20	2
merveilleux	6	1	5
minuscule	1	1	0
mirobolant	0	0	0
moderne	9	0	9
modeste	3	0	3
moindre	7	7	0
moyen	2	1	1
multiple	2	1	1
mystérieux	1	0	1
nécessaire	2	0	2
net	5	2	3
nombreux	17	16	1
nouveau	39	31	8
périlleux	2	0	2
pertinent	0	0	0
pesant	0	0	0
petit	100	99	1
plein	15	13	2
possible	7	1	6
précédent	6	1	5
précieux	0	0	0
premier	100	100	0
présent	2	0	2
prestigieux	0	0	0
prétendu	0	0	0
principal	17	11	6
probable	1	1	0

Adjectif	total	ant	post
prochain	15	9	6
proche	0	0	0
profond	6	2	4
propre	30	26	4
prudent	0	0	0
puissant	4	1	3
pur	9	2	7
quelconque	5	0	5
rapide	4	0	4
rare	2	2	0
récent	4	2	2
redoutable	0	0	0
réel	6	1	5
regrettable	0	0	0
relatif	4	1	3
remarquable	1	0	1
rentable	2	0	2
riche	3	2	1
rigoureux	2	0	2
sacro-saint	0	0	0
sage	0	0	0
sain	1	0	1
salutaire	0	0	0
sensible	0	0	0
sérieux	5	0	5
seul	38	37	1
sévère	1	0	1
simple	12	3	9
solide	1	0	1
sombre	2	0	2
soudain	1	1	0
sournois	0	0	0
strict	1	0	1
substantiel	0	0	0
total	12	2	10
traditionnel	4	0	4
ultime	0	0	0
unique	2	0	2
utile	0	0	0
véritable	14	13	1
vif	4	2	2
vigoureux	0	0	0
violent	5	1	4
vrai	23	21	2

E.3.1.4. Wilmet (1980)

Adjectif	total	ant	post
actuel	17	4	13
aléatoire	0	0	0
ambitieux	4	2	2
amer	29	4	25
ample	14	13	1
ancien	132	89	43
apparent	18	2	16
autre	68	65	3
banal	17	5	12
bas	116	32	84
bon	479	467	12
bref	43	26	17
britannique	2	0	2
brusque	42	18	24
brutal	22	2	20
célèbre	12	6	6
certain	46	41	5
chaud	73	11	62
classique	55	1	54
complet	37	3	34
confortable	12	4	8
considérable	9	0	9
constant	18	8	10
contraignant	1	0	1
court	96	43	53
coûteux	2	0	2
dangereux	16	6	10
délicat	39	16	23
dernier	116	88	28
différent	39	9	30
difficile	40	3	37
dit	7	0	7
divers	30	13	17
douloureux	23	3	20
dramatique	0	0	0
dur	82	14	68
écrasant	5	1	4
égal	21	3	18

Adjectif	total	ant	post
éminent	2	2	0
énorme	62	37	25
esthétique	11	0	11
étrange	78	48	30
étroit	54	15	39
éventuel	3	0	3
exact	20	2	18
excellent	38	34	4
exceptionnel	17	0	17
extraordinaire	28	12	16
extrême	31	16	15
faible	33	26	7
faux	50	40	10
ferme	18	3	15
fidèle	14	4	10
fin	51	19	32
flagrant	1	1	0
flamboyant	4	0	4
formidable	14	6	8
fort	79	33	46
fou	42	14	28
franc	15	6	9
futur	25	9	16
gigantesque	12	10	2
grand	1304	1262	42
grave	40	10	30
habituel	32	5	27
haut	148	113	35
heureux	46	10	36
immédiat	22	0	22
important	21	4	17
impressionnant	4	0	4
indispensable	6	1	5
inégal	15	1	14
inéluctable	6	3	3
inévitable	6	0	6
inexorable	3	0	2
influent	0	0	0

Adjectif	total	ant	post
inquiétant	2	0	2
insuffisant	1	0	1
intense	11	3	8
interminable	27	14	13
irrésistible	9	4	5
irréversible	3	0	3
jeune	452	424	28
juste	11	10	1
large	62	31	31
légendaire	2	0	2
léger	130	77	53
lent	44	13	31
libre	49	8	41
lointain	55	17	38
long	300	238	62
louable	2	2	0
lourd	81	37	44
luxueux	1	0	1
majeur	10	3	7
malheureux	29	14	15
mauvais	152	145	7
médiocre	14	7	7
meilleur	30	28	2
merveilleux	36	18	18
minuscule	32	17	15
mirobolant	0	0	0
moderne	25	2	23
modeste	14	4	10
moindre	64	63	1
moyen	0	0	0
multiple	9	3	6
mystérieux	48	18	30
nécessaire	26	1	25
net	22	1	21
nombreux	22	10	12
nouveau	221	141	80
périlleux	3	1	2
pertinent	2	0	2
pesant	9	2	7
petit	1139	1124	15
plein	155	79	76
possible	28	1	27
précédent	15	5	10
précieux	31	11	20
premier	0	0	0
présent	18	3	15
prestigieux	4	0	4

Adjectif	total	ant	post
prétendu	0	0	0
principal	31	11	20
probable	5	0	5
prochain	46	26	20
proche	40	8	32
profond	102	35	67
propre	138	114	24
prudent	3	0	3
puissant	41	13	28
pur	87	27	60
quelconque	11	7	4
rapide	49	9	40
rare	40	21	19
récent	13	6	7
redoutable	9	4	5
réel	25	2	23
regrettable	2	0	2
relatif	9	0	9
remarquable	8	2	6
rentable	2	0	2
riche	23	10	13
rigoureux	13	2	11
sacro-saint	0	0	0
sage	16	8	8
sain	9	2	7
salutaire	2	0	2
sensible	26	3	23
sérieux	27	1	26
seul	247	210	37
sévère	18	1	17
simple	94	50	44
solide	29	11	18
sombre	83	17	66
soudain	21	7	14
sournois	19	2	17
strict	6	2	4
substantiel	2	0	2
total	27	3	24
traditionnel	8	1	7
ultime	9	6	3
unique	66	35	31
utile	8	2	6
véritable	66	44	22
vif	67	23	44
vigoureux	5	0	5
violent	44	15	29
vrai	142	130	12

E.3.2. Proportions d'antéposition des adjectifs alternant dans les quatre corpus

Dans ce tableau, lorsqu'une case est vide, cela signifie qu'il n'y a pas de données pour l'adjectif dans le corpus concerné.

Adjectif	FTB	ESTER	CORALROM	Wilmet
actuel	18.40	31.00	10.00	23.50
aléatoire	16.70	0.00		
ambitieux	37.50		100.00	50.00
amer	25.00	0.00	0.00	13.80
ample	50.00			92.90
ancien	46.70	90.90	78.90	67.40
apparent	50.00	50.00	50.00	11.10
autre	99.50	99.00	99.00	95.60
banal	40.00	0.00	0.00	29.40
bas	70.20	77.80	50.00	27.60
bon	99.20	98.00	100.00	97.50
bref	60.00	100.00	100.00	60.50
britannique	1.10	0.00	0.00	0.00
brusque	50.00	100.00		42.90
brutal	28.60	0.00	0.00	9.10
célèbre	87.50	73.30	50.00	50.00
certain	92.20	100.00	97.30	89.10
chaud	33.30	11.10		15.10
classique	12.50	8.30	0.00	1.80
complet	5.90	7.70	0.00	8.10
confortable	50.00	100.00		33.30
considérable	6.70	11.10	0.00	0.00
constant	16.70	25.00	0.00	44.40
contraignant	33.30			0.00
court	69.70	43.80	100.00	44.80
coûteux	25.00	100.00	0.00	0.00
dangereux	50.00	25.00	0.00	37.50
délicat	23.10	33.30	0.00	41.00
dernier	61.10	78.00	60.30	75.90
différent	58.80	45.20	33.30	23.10
difficile	25.00	5.30	0.00	7.50
dit	50.00	0.00		0.00
divers	21.40	14.30	50.00	43.30
douloureux	20.00	0.00	100.00	13.00
dramatique	20.00		0.00	
dur	46.20	0.00	0.00	17.10

Adjectif	FTB	ESTER	CORALROM	Wilmet
écrasant	50.00	0.00	0.00	20.00
égal	33.30	0.00	0.00	14.30
éminent	50.00	0.00	100.00	100.00
énorme	57.90	100.00	26.70	59.70
esthétique	25.00		0.00	0.00
étrange	75.00	66.70	25.00	61.50
étroit	8.30	20.00	100.00	27.80
éventuel	82.10	78.30	0.00	0.00
exact	20.00	0.00	0.00	10.00
excellent	77.80	80.00	50.00	89.50
exceptionnel	6.90	0.00	0.00	0.00
extraordinaire	10.00	11.10	0.00	42.90
extrême	80.00	42.90	50.00	51.60
faible	66.70	50.00	33.30	78.80
faux	83.30	100.00	100.00	80.00
ferme	40.00	0.00		16.70
fidèle	25.00	66.70	0.00	28.60
fin	50.00	25.00	0.00	37.30
flagrant	50.00		0.00	100.00
flamboyant	50.00		0.00	0.00
formidable	84.60	100.00	100.00	42.90
fort	61.50	58.30	33.30	41.80
fou	33.30	50.00	0.00	33.30
franc	33.30	50.00	0.00	40.00
futur	80.00	97.20	66.70	36.00
gigantesque	80.00	66.70	0.00	83.30
grand	99.70	97.00	95.00	96.80
grave	72.20	26.30	0.00	25.00
habituel	7.70	0.00	0.00	15.60
haut	96.30	100.00	76.90	76.40
heureux	50.00	0.00	50.00	21.70
immédiat	9.10	0.00	0.00	0.00
important	42.50	15.50	13.60	19.00
impressionnant	50.00	0.00	0.00	0.00
indispensable	57.10	33.30	0.00	16.70
inégal	25.00			6.70
inéluctable	20.00		0.00	50.00
inévitable	80.00	20.00		0.00
inexorable	50.00		0.00	0.00
influent	50.00	0.00		
inquiétant	28.60	28.60		0.00
insuffisant	16.70			0.00
intense	66.70	66.70	0.00	27.30

Adjectif	FTB	ESTER	CORALROM	Wilmet
interminable	50.00			51.90
irrésistible	25.00	0.00		44.40
irréversible	50.00			0.00
jeune	89.50	96.60	93.80	93.80
juste	87.50	54.50	100.00	90.90
large	61.90	80.00	0.00	50.00
légendaire	50.00			0.00
léger	81.20	66.70	87.50	59.20
lent	42.90	0.00	0.00	29.50
libre	47.40	18.80	50.00	16.30
lointain	50.00	0.00		30.90
long	73.80	77.80	68.40	79.30
louable	33.30			100.00
lourd	38.90	17.60	20.00	45.70
luxueux	50.00	50.00		0.00
majeur	18.20	9.10	11.10	30.00
malheureux	33.30	0.00	50.00	48.30
mauvais	97.70	100.00	100.00	95.40
médiocre	33.30	0.00		50.00
meilleur	91.50	97.30	90.90	93.30
merveilleux	50.00	0.00	16.70	50.00
minuscule	50.00	50.00	100.00	53.10
mirobolant	50.00	0.00		
moderne	11.10	0.00	0.00	8.00
modeste	18.20	28.60	0.00	28.60
moindre	94.30	100.00	100.00	98.40
moyen	11.90	42.90	50.00	
multiple	76.50	50.00	50.00	33.30
mystérieux	50.00	33.30	0.00	37.50
nécessaire	17.60	16.70	0.00	3.80
net	22.20	50.00	40.00	4.50
nombreux	94.10	93.00	94.10	45.50
nouveau	87.50	96.00	79.50	63.80
périlleux	33.30	0.00	0.00	33.30
pertinent	50.00			0.00
pesant	50.00	0.00		22.20
petit	97.10	100.00	99.00	98.70
plein	85.00	82.60	86.70	51.00
possible	12.50	25.00	14.30	3.60
précédent	17.20	17.60	16.70	33.30
précieux	50.00	0.00		35.50
premier	99.50	100.00	100.00	
présent	10.00	0.00	0.00	16.70
prestigieux	28.60	100.00		0.00

Adjectif	FTB	ESTER	CORALROM	Wilmet
prétendu	87.50			
principal	83.10	75.90	64.70	35.50
probable	66.70	16.70	100.00	0.00
prochain	59.00	45.50	60.00	56.50
proche	50.00	33.30		20.00
profond	37.50	10.00	33.30	34.30
propre	80.40	90.00	86.70	82.60
prudent	33.30	0.00		0.00
puissant	54.50	50.00	25.00	31.70
pur	50.00	0.00	22.20	31.00
quelconque	50.00	100.00	0.00	63.60
rapide	5.70	7.70	0.00	18.40
rare	75.00	72.70	100.00	52.50
récent	40.00	25.00	50.00	46.20
redoutable	75.00	33.30		44.40
réel	23.10	29.40	16.70	8.00
regrettable	50.00	0.00		0.00
relatif	50.00	0.00	25.00	0.00
remarquable	50.00	40.00	0.00	25.00
rentable	25.00		0.00	0.00
riche	10.00	33.30	66.70	43.50
rigoureux	12.50	50.00	0.00	15.40
sacro-saint	80.00			
sage	50.00			50.00
sain	40.00		0.00	22.20
salutaire	50.00			0.00
sensible	11.50	0.00		11.50
sérieux	59.40	16.70	0.00	3.70
seul	92.90	100.00	97.40	85.00
sévère	42.90	0.00	0.00	5.60
simple	92.60	71.90	25.00	53.20
solide	46.70	0.00	0.00	37.90
sombre	50.00	0.00	0.00	20.50
soudain	33.30		100.00	33.30
sournois	50.00			10.50
strict	50.00	0.00	0.00	33.30
substantiel	16.70			0.00
total	4.10	11.50	16.70	11.10
traditionnel	30.30	22.20	0.00	12.50
ultime	88.90	100.00		66.70
unique	21.70	23.50	0.00	53.00
utile	16.70	0.00		25.00
véritable	96.80	100.00	92.90	66.70
vif	73.10	50.00	50.00	34.30
vigoureux	42.90	0.00		0.00
violent	42.90	33.30	20.00	34.10
vrai	80.00	100.00	91.30	91.50

Annexe F

Intercepts aléatoires relatifs aux verbes

Cette annexe présente la liste des intercepts aléatoires associés à la variable `lemSem` dans le Modèle *TF* qui est reproduit dans la table F.1. Les intercepts aléatoires sont accompagnés du nombre d'occurrences dans la table de données *TF*, ainsi que des bornes supérieures et inférieures de l'intervalle de confiance à 95%. Les intercepts aléatoires sont présentés par classe sémantique et rangés dans l'ordre croissant.

```
Effets aléatoires :
  Groupes  Nom          Variance  Ecart-type
  lemSem   (Intercept)  1.24298   1.11489
  corpus   (Intercept)  0.24245   0.49239
Nombre d'obs. : 1434 ; groupes : lemSem, 253 ; corpus, 4

Effets fixes :
               Estimation  Erreur-type  valeur z  Pr(>|z|)
  (Intercept)  -1.4269     0.2879       -4.955    7.22e-07  ***
  longRelMots  2.6891      0.1565       17.183    < 2e-16   ***

Corrélation des effets fixes :
              (Intercept)
  longRelMots -0.128
```

Table F.1.: Les paramètres du Modèle *TF*

Communication

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
annoncer C	-0.641724171	-1.483493860	0.200045518	3
interdire C	-0.404546661	-1.327817987	0.518724665	3
informer C	-0.381185801	-1.341512526	0.579140924	4
imposer C	-0.369628592	-1.357149766	0.617892582	1
faire C	-0.353060278	-1.288562396	0.582441839	8
transmettre C	-0.254613864	-1.083232940	0.574005213	4
accuser C	-0.233840240	-1.275685656	0.808005175	2
porter C	-0.228229968	-1.276454370	0.819994435	1
permettre C	-0.171828302	-1.123179859	0.779523256	3
donner C	-0.157798743	-1.243772915	0.928175430	5
fixer C	-0.153532292	-0.818269387	0.511204802	8
manifester C	-0.096298781	-1.238091553	1.045493991	1
convaincre C	-0.059188444	-1.233960851	1.115583962	5
refuser C	-0.013593457	-1.240002073	1.212815158	2
remercier C	-0.008730572	-1.241031033	1.223569889	1
confier C	-0.000952371	-1.242751688	1.240846946	1
prévenir C	0.014085105	-1.046078713	1.074248922	5
rappeler C	0.060959936	-0.677387324	0.799307195	3
expliquer C	0.147629476	-0.703944269	0.999203221	5
suggérer C	0.155634277	-0.932925536	1.244194090	2
adresser C	0.157992825	-0.837560709	1.153546359	3
demander C	0.230851796	-0.049281823	0.510985415	22
livrer C	0.277274032	-0.723712556	1.278260620	2
communiquer C	0.330434292	-0.633519218	1.294387801	3
négocier C	0.352664841	-0.641034795	1.346364476	1
présenter C	0.354769330	-0.363512712	1.073051371	7
montrer C	0.392100207	-0.056177265	0.840377678	12
opposer C	0.398398866	-0.556649625	1.353447357	2
appeler C	0.464344657	-0.302936179	1.231625494	5
dicter C	0.521238019	-0.198467231	1.240943269	5
notifier C	0.600723295	-0.115262319	1.316708910	3
confirmer C	0.745748810	-0.044877252	1.536374872	4
exiger C	0.766499931	0.166654620	1.366345241	4
préconiser C	0.929562409	0.176307904	1.682816914	4
apprendre C	1.150110185	0.522627353	1.777593016	4
proposer C	1.184045730	0.760584317	1.607507143	10

Don, privation

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
verser D	-1.121729971	-1.747543535	-0.495916407	11
ouvrir D	-1.049396180	-1.562611216	-0.536181145	20
deleguer D	-0.871498222	-1.738136347	-0.004860096	3
porter D	-0.842444022	-1.620676068	-0.064211977	9
apporter D	-0.833651747	-1.246617020	-0.420686473	19
abandonner D	-0.747663269	-1.599062724	0.103736186	3
céder D	-0.662320539	-1.177021509	-0.147619569	21
accorder D	-0.661442057	-1.082246801	-0.240637312	17
réserver D	-0.634067380	-1.291750173	0.023615413	7
interdire D	-0.622174104	-1.561322258	0.316974051	2
affecter D	-0.610088933	-1.560375672	0.340197805	1
souffler D	-0.600602672	-1.456573541	0.255368197	4
garantir D	-0.558955007	-1.036083725	-0.081826289	8
confier D	-0.468865585	-1.027710461	0.089979291	8
infliger D	-0.446632058	-1.247876233	0.354612118	4
rapporter D	-0.374537882	-0.948493700	0.199417936	5
appliquer D	-0.343526593	-1.199331452	0.512278266	5
communiquer D	-0.336725920	-1.334338272	0.660886433	2
voler D	-0.333236967	-0.992278209	0.325804275	4
transmettre D	-0.304611758	-1.026503250	0.417279735	8
payer D	-0.258351300	-0.756787027	0.240084426	12
intéresser D	-0.236390071	-1.277043170	0.804263028	3
distribuer D	-0.234508988	-0.763050566	0.294032590	9
ajouter D	-0.224057552	-1.274984571	0.826869467	1
garder D	-0.224057552	-1.274984571	0.826869467	1
attribuer D	-0.173307267	-0.635764641	0.289150108	12
consacrer D	-0.112367213	-0.845113938	0.620379512	10
appuyer D	-0.084932888	-1.236871210	1.067005434	1
opposer D	-0.077205140	-1.236345438	1.081935158	1
passer D	-0.077171253	-1.233376694	1.079034189	7
décerner D	-0.068598607	-0.761108116	0.623910903	4
arracher D	-0.049745878	-1.236113061	1.136621306	1
coûter D	-0.049745878	-1.236113061	1.136621306	1
adresser D	-0.048391044	-1.236240002	1.139457914	1
assurer D	-0.035653677	-0.540493752	0.469186399	6
faire D	-0.028388181	-1.237833549	1.181057187	1
retirer D	-0.028388181	-1.237833549	1.181057187	1
racheter D	-0.024944378	-0.675924896	0.626036139	5
léguer D	-0.019922956	-0.727944192	0.688098280	5
mettre D	0.052458666	-0.762738719	0.867656050	4

Don, privation (suite)

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
remettre D	0.108324228	-0.414241779	0.630890235	8
obtenir D	0.110977919	-0.408319110	0.630274947	7
transférer D	0.171112385	-0.912604283	1.254829054	1
redonner D	0.171631895	-0.169363626	0.512627416	19
refuser D	0.217955591	-0.756333710	1.192244892	2
suggérer D	0.238368109	-0.804320111	1.281056329	1
reprendre D	0.242710356	-0.798072640	1.283493353	1
laisser D	0.276937624	-0.100975588	0.654850835	18
permettre D	0.358095776	-0.595919636	1.312111188	2
soumettre D	0.388799116	-0.345848442	1.123446674	6
donner D	0.428282457	0.328045735	0.528519180	91
octroyer D	0.482704666	-0.316501912	1.281911244	3
fournir D	0.488262464	-0.073455665	1.049980593	10
concéder D	0.507282384	-0.290544775	1.305109543	3
attirer D	0.516064072	-0.439764974	1.471893119	1
offrir D	0.523606950	0.187801402	0.859412498	16
devoir D	0.530882458	-0.240099566	1.301864483	7
livrer D	0.636654051	-0.019416732	1.292724833	5
ramener D	0.771467656	-0.191076043	1.734011355	1
retarder D	0.994417088	-0.043322984	2.032157160	1
imposer D	1.035190083	0.562013071	1.508367094	14
conférer D	1.263459905	0.696879271	1.830040540	5
recevoir D	1.275278280	0.591581206	1.958975355	5
vendre D	1.356915555	1.119538868	1.594292242	30
acheter D	1.621497097	0.922719085	2.320275110	7
rendre D	1.715727764	1.325666587	2.105788941	13

Entrée, sortie

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
renvoyer E	-1.310327975	-2.113327219	-0.507328731	7
expédier E	-1.134298764	-1.895650858	-0.372946670	4
passer E	-1.042506161	-1.759248343	-0.325763979	8
élargir E	-0.769508840	-1.654224437	0.115206757	6
conduire E	-0.633917714	-1.567324531	0.299489103	3
transférer E	-0.557044675	-1.233072744	0.118983394	7
déverser E	-0.490995534	-1.293621342	0.311630274	3
tirer E	-0.432952236	-1.022012220	0.156107748	8
apporter E	-0.430732817	-1.403296585	0.541830952	1
attirer E	-0.417263061	-1.365719758	0.531193637	3
rapprocher E	-0.402537504	-1.325265136	0.520190129	4
amener E	-0.380713881	-1.084996824	0.323569063	6
sortir E	-0.356151005	-1.021209534	0.308907523	10
sauver E	-0.306802949	-1.110178695	0.496572797	6
concentrer E	-0.245260679	-1.277835203	0.787313845	4
rapporter E	-0.238216760	-1.004937893	0.528504374	3
exporter E	-0.210403484	-1.000528529	0.579721561	3
orienter E	-0.153497988	-1.238969259	0.931973283	4
nommer E	-0.143743903	-1.246969597	0.959481791	1
réorienter E	-0.114878413	-1.233715332	1.003958506	4
refuser E	-0.098500246	-1.238254109	1.041253616	1
retenir E	-0.096298781	-1.238091553	1.045493991	1
accompagner E	-0.048541284	-1.236194192	1.139111624	1
assigner E	-0.048541284	-1.236194192	1.139111624	1
placer E	-0.048391044	-1.236240002	1.139457914	1
adresser E	-0.031921749	-0.799423017	0.735579519	4
introduire E	-0.017727092	-1.239366541	1.203912357	1
porter E	0.029219247	-0.278136836	0.336575329	28
déplacer E	0.054023874	-1.127963343	1.236011091	1
réduire E	0.199334145	-0.541550192	0.940218481	7
retirer E	0.260813844	-0.541342052	1.062969739	2
ramener E	0.346552555	-0.061629746	0.754734857	21
retarder E	0.352664841	-0.641034795	1.346364476	1
appeler E	0.468062254	-0.384509424	1.320633933	3
écarter E	0.477105655	-0.308714980	1.262926291	3
déduire E	0.585842304	-0.081816388	1.253500996	6
arracher E	0.590387455	-0.359533507	1.540308418	1
dépasser E	0.895129583	0.030237521	1.760021645	2
exclure E	1.283277029	0.539341226	2.027212832	4

Frapper, toucher

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
soupçonner F	-0.610088933	-1.560375672	0.340197805	1
donner F	-0.230648641	-1.275695441	0.814398160	2
bombarder F	-0.096298781	-1.238091553	1.045493991	1
accuser F	-0.022069422	-1.238453639	1.194314795	2
menacer F	0.043703264	-0.737362439	0.824768967	4
qualifier F	0.352664841	-0.641034795	1.346364476	1

États physiques et comportements

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
évaluer H	-0.919440997	-1.600705017	-0.238176977	3
payer H	-0.437363110	-1.406737292	0.532011073	1
inciter H	-0.407140259	-1.312957453	0.498676934	7
forcer H	-0.302452595	-1.289763058	0.684857868	3
conduire H	-0.111153827	-1.237594973	1.015287319	2
plier H	-0.096298781	-1.238091553	1.045493991	1
contraindre H	-0.086235568	-0.759842444	0.587371308	4
préparer H	-0.075418826	-1.236317962	1.085480310	1
prendre H	-0.048793224	-0.712526648	0.614940200	6
amener H	-0.013067434	-1.240190600	1.214055732	1
coûter H	0.009640217	-0.858509363	0.877789797	2
chiffrer H	0.245782776	-0.527247980	1.018813532	4
estimer H	1.342724896	0.611314831	2.074134961	4

Mouvement sur place

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
augmenter M	-0.613079092	-1.097554264	-0.128603919	12
terminer M	-0.525549366	-1.481080247	0.429981515	1
conclure M	-0.105127069	-1.233778653	1.023524514	3
limiter M	-0.017727092	-1.239366541	1.203912357	1
étendre M	0.025401992	-1.187421080	1.238225064	1
accroître M	0.226678674	-0.575661417	1.029018765	5
abaisser M	0.449840413	-0.420494404	1.320175231	3
réduire M	0.799109082	0.484033155	1.114185010	13
diminuer M	2.556739130	2.039450874	3.074027387	11

Locatif

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
laisser L	-1.134270621	-1.850873476	-0.417667765	6
indexer L	-0.893105324	-1.680715060	-0.105495587	4
mettre L	-0.746406278	-0.988266439	-0.504546118	51
prendre L	-0.649942226	-1.496782052	0.196897600	4
aligner L	-0.610088933	-1.560375672	0.340197805	1
porter L	-0.406814927	-1.350329410	0.536699555	2
déposer L	-0.343016152	-0.976037759	0.290005456	9
placer L	-0.339444705	-0.844663590	0.165774181	13
limiter L	-0.247346884	-0.629648205	0.134954436	13
introduire L	-0.245370472	-1.261120351	0.770379406	3
appuyer L	-0.145359787	-1.246804886	0.956085313	3
fonder L	-0.130881585	-1.237963570	0.976200400	5
baser L	-0.025481564	-1.237830511	1.186867383	3
classer L	-0.008102827	-1.241159943	1.224954288	1
parier L	-0.003470294	-1.242165113	1.235224526	1
inscrire L	0.101122708	-0.575777454	0.778022869	6
remettre L	0.278058088	-0.195358546	0.751474723	13
trouver L	0.696402862	0.438162779	0.954642946	20
installer L	0.884563285	0.236471159	1.532655411	8

Munir, démunir

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
équiper N	-0.541569813	-1.212045123	0.128905498	4
protéger N	-0.304388076	-1.275608159	0.666832008	5
doter N	-0.220062943	-1.250977275	0.810851390	4
ceindre N	-0.096298781	-1.238091553	1.045493991	1
vider N	0.375467447	-0.465974688	1.216909582	4
bourrer N	0.905128238	-0.093757140	1.904013615	1

Verbes psychologiques

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
préférer P	-1.135561053	-1.798225814	-0.472896291	7
justifier P	-0.426844714	-1.343017731	0.489328303	7
porter P	-0.369021935	-1.319192251	0.581148382	4
apporter P	-0.224316700	-1.275000797	0.826367396	2
consacrer P	-0.114622401	-1.234422368	1.005177567	4
trouver P	-0.027058854	-1.238037220	1.183919512	1
limiter P	-0.002823993	-1.242312834	1.236664848	1
inspirer P	0.372834261	-0.494209720	1.239878242	3

Réalisation, mise en état

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
faire R	-0.590023579	-1.114186797	-0.065860360	19
évaluer R	-0.106244762	-0.690645410	0.478155887	6
lancer R	-0.027675449	-1.237932105	1.182581208	1
conclure R	0.043033724	-0.813776732	0.899844180	2
mettre R	0.465544855	0.117591163	0.813498546	19
remettre R	0.493361040	-0.306921972	1.293644053	3
appliquer R	0.631511336	-0.317908295	1.580930967	1

Saisir, serrer, posséder

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
priver S	-0.919693098	-1.713342210	-0.126043985	5
recevoir S	-0.610088933	-1.560375672	0.340197805	1
voler S	-0.437363110	-1.406737292	0.532011073	1
tirer S	-0.280730154	-1.284348685	0.722888377	2
trouver S	-0.160149600	-1.240088796	0.919789596	4
arracher S	-0.136172322	-1.235184864	0.962840220	3
laisser S	0.059390075	-0.557112677	0.675892827	9
retirer S	0.274210936	-0.750953613	1.299375486	1
prendre S	0.361184646	-0.590419136	1.312788429	3
ôter S	0.403542586	-0.318120977	1.125206148	5
emprunter S	0.724601825	0.003311123	1.445892527	5
obtenir S	1.471991881	0.563862723	2.380121039	4

Transformation, changement

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
remplacer T	-1.108128448	-1.680796145	-0.535460751	12
echanger T	-0.502629863	-1.383116169	0.377856443	5
convertir T	-0.270005190	-1.095122503	0.555112124	3
transformer T	-0.226650783	-0.876139212	0.422837646	8
ériger T	-0.224057552	-1.274984571	0.826869467	1
compléter T	-0.088408686	-1.235520012	1.058702640	4
troquer T	0.370227431	-0.362407274	1.102862135	5
faire T	3.268607650	2.972838676	3.564376625	31

Union, réunion

	(Intercept)	Borne inférieure	Borne supérieure	Nombre d' occurrences
séparer U	-1.288112989	-1.983828954	-0.592397024	7
comparer U	-0.624768805	-1.436436862	0.186899252	6
distinguer U	-0.599621936	-1.426371842	0.227127970	5
assimiler U	-0.548814115	-1.405076353	0.307448123	5
allier U	-0.518584672	-1.371229802	0.334060457	5
lier U	-0.424180926	-1.364651926	0.516290075	4
opposer U	-0.420183648	-1.024766573	0.184399277	11
affranchir U	-0.413760139	-1.334492289	0.506972012	4
protéger U	-0.228229968	-1.276454370	0.819994435	1
concilier U	-0.224057552	-1.274984571	0.826869467	1
laver U	-0.224057552	-1.274984571	0.826869467	1
concentrer U	-0.159780492	-1.251122749	0.931561765	1
adapter U	-0.154779179	-1.239638469	0.930080111	4
exonérer U	-0.140915380	-1.239186506	0.957355745	3
intéresser U	-0.011531494	-1.240476827	1.217413838	1
ouvrir U	-0.001081063	-1.242721026	1.240558900	1
adjoindre U	0.018271989	-1.202737575	1.239281553	1
verser U	0.352664841	-0.641034795	1.346364476	1
associer U	0.435163380	-0.150765971	1.021092732	9
priver U	0.611538572	-0.319758040	1.542835184	3
rallier U	0.663438230	-0.164537337	1.491413797	4
écarter U	1.694720527	0.882782091	2.506658963	3
ajouter U	2.023583555	1.404110784	2.643056326	9

Bibliographie

Abeillé A. à paraître. La place de l'adjectif épithète. In A. Abeillé, D. Godard & A. Delaveau, Eds., *La grande grammaire du français*. Paris : Bayard.

Abeillé A. & Barrier N. 2004. Enriching a French treebank. In *Proceedings of Language Ressources and Evaluation Conference (*Lrec*)*, Lisbonne.

Abeillé A., Clément L. & Toussenel F. 2003. Building a treebank for French. In *Treebanks*. Dordrecht : Kluwer.

Abeillé A. & Godard D. 1999. La position de l'adjectif épithète en français : le poids des mots. *Recherches Linguistiques de Vincennes*, **28**, 9–32.

Abeillé A. & Godard D. 2000. French word order and lexical weigh. In R. Borsley, Ed., *The Nature and Function of Syntactic Categories*, volume 32 of *Syntax and Semantics*, p. 325–358. New York : Academic Press.

Abeillé A. & Godard D. 2001. A class of lite adverbs in French. In J. Camps & C. Wiltshire, Eds., *Romance Syntax, Semantics and their L2 Acquisition*, p. 9–25. Amsterdam : John Benjamins.

Abeillé A. & Godard D. 2004. De la légèreté en syntaxe. *Bulletin de la Société de Linguistique de Paris*, **XCIX**(1), 69–106.

Abeillé A. & Godard D. 2006. La légèreté en français comme déficience de mobilité. *Lingvisticae Investigationes*, **29**(1), 11–24.

Ågren J. 1973. *Enquête sur quelques liaisons facultatives dans le français de conversation radiophonique*. Upsala : Acta Universitatis Upsaliensis.

Agresti A. 2007. *An Introduction to Categorical Data Analysis*. Wiley.

Allan K. 1987. Hierarchies and the choice of left conjuncts. *Journal of Linguistics*, **23**, 51–77.

Ariel M. 1990. *Accessing Noun-Phrase Antecedents*. London : Routledge.

Arnold J. E. 1998. *Reference Form and Discourse Patterns*. PhD thesis, Stanford University, Stanford.

Arnold J. E., Wasow T., Losongco A. & Ginstrom R. 2000. Heaviness vs. newness : The effects of structural complexity and discourse status on constituent ordering. *Language*, **76**(1), 28–55.

Artstein R. & Poesio M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**, 555–596.

Baayen H. R. 2008. *Analyzing linguistic data : A practical introduction to statistics using R*. New York : Cambridge University Press.

Baayen R. H. 2003. Probabilistic approaches to morphology. In R. Bod, J. Hay & S. Jannedy, Eds., *Probabilistic Linguistics*, p. 229–287. Cambridge, MA : The MIT Press.

Baayen R. H., Davidson D. & Bates D. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, **59**(4), 390–412.

Baayen R. H., Feldman L. B. & Schreuder R. 2006. Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, **55**, 290–313.

Bader M. & Häussler J. 2010. Word order in German : A corpus study. *Lingua*, **120**, 717–762.

Baroni M. & Evert S. 2008. Statistical methods for corpus exploitation. In A. Lüeling & M. Kytö, Eds., *Corpus Linguistics. An International Handbook*, p. 777–802, Berlin : Mouton de Gruyter.

Bartning I. 1976. *Remarques sur la syntaxe et la sémantique des pseudo-adjectifs dénominaux en français*. PhD thesis, Université de Stockholm.

Bates D. & Sarkar D. 2007. lme4 : Linear mixed-effects models using S4 classes. R package version 0.9975-13.

Beckner C. & Bybee J. 2009. A usage-based account of constituency and reanalysis. *Language Learning*, **59**, 27–46.

Behaghel O. 1909. Von deutscher Wortstellung [de l'ordre des mots en allemand]. *Indogermanische Forschungen*, **25**, 110–142.

Behaghel O. 1932. *Deutsche Syntax : eine geschichtliche Darstellung, Band IV, Wortstellung*. Heidelberg : C. Winter.

Belsley D. A., Kuh E. & Welsch R. E. 1980. *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*. New York : Wiley.

Benor S. B. & Levy R. 2006. The chicken or the egg ? A probabilistic analysis of English binomials. *Language*, **82**(2), 28–55.

Berrendonner A. 1987. L'ordre des mots et ses fonctions. *Travaux de linguistique*, **14/15**, 9–19.

Beyssade C., Delais-Roussarie E., Doetjes J., Marandin J.-M. & Rialland A. 2004a. Prosodic, syntactic and pragmatic aspects of information structure − An introduction. In F. Corblin & H. de Swart, Eds., *Handbook of French Semantics*, p. 455–473. CSLI Publications.

Beyssade C., Delais-Roussarie E., Doetjes J., Marandin J.-M. & Rialland A. 2004b. Prosody and information in French. In F. Corblin & H. de Swart, Eds., *Handbook of French Semantics*, p. 455–473. CSLI Publications.

Blache P. 2005. Property grammars : A fully constraint-based theory. In H. Christiansen, P. R. Skadhauge & J. Villadsen, Eds., *Constraint Solving and Language Processing*. Springer.

Blache P. 2010. Un modèle de caractérisation de la complexité syntaxique. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal.

Blache P., Hemforth B. & Rauzy S. 2006. Acceptability prediction by means of grammaticality quantification. In N. Calzolari, C. Cardie & P. Isabelle, Eds., *Proceedings of ACL 2006* : The Association for Computer Linguistics.

Blinkenberg A. 1928. *L'ordre des mots en français moderne. Première partie.* Copenhague : Høst & Søn.

Blinkenberg A. 1933. *L'ordre des mots en français moderne. Deuxième partie.* Copenhague : Levin & Munksgaard.

Bock J. K. & Warren R. K. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, **21**, 47–67.

Bock K. J. 1986. Meaning, sound, and syntax : Lexical priming in sentence production. *Journal of Experimental Psychology : Learning, Memory and Cognition*, **124**, 575–586.

Bock K. J. 1987. An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, **26**, 119–137.

Bock K. J. & Irwin D. E. 1980. Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior*, **19**, 467–484.

Bock K. J., Loebell H. & Morey R. 1992. From conceptual roles to structural relations : Bridging the syntactic cleft. *Psychological Review*, **99**(1), 150–171.

Boersma P. 1998. *Functional Phonology. Formalizing the interaction between articulatory and perceptual drives*. PhD thesis, University of Amsterdam, Amsterdam.

Boersma P. 2000. Learning a grammar in Functional Phonology. In J. Dekkers, F. van der Leeuw & J. van de Weijer, Eds., *Optimality Theory : Phonology, Syntax, and Acquisition*, p. 465–523. Oxford University Press.

Boersma P. & Hayes B. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, **32**, 45–86.

Bonami O. & Boyé G. 2003. La nature morphologique des allomorphies conditionnées : les formes de liaison des adjectifs en français. In *Actes du troisième forum de morphologie*, Collection Silexicales, Université Lille 3.

Bonami O. & Boyé G. 2005. Construire le paradigme d'un adjectif. *Recherches Linguistiques de Vincennes*, **34**, 77–98.

Bonami O. & Delais-Roussarie E. à paraître. Phénomènes segmentaux. In A. Abeillé, D. Godard & A. Delaveau, Eds., *La grande grammaire du français*. Paris : Bayard.

Bouchard D. 1998. The distribution and interpretation of adjectives in French : A consequence of bare phrase structure. *Probus*, **10**(2), 139–183.

Branigan H., Pickering M. & Cleland A. 1999. Syntactic priming in written production : Evidence for rapid decay. *Psychonomic Bulletin and Review*, **6**, 635–640.

Branigan H. P. & Feleki E. 1999. Conceptual accessibility and serial order in Greek language production. In M. Hahn & S. C. Stoness, Eds., *Proceedings of the 21st Conference of the Cognitive Science Society*, p. 96–101, Mahwah : Erlbaum.

Branigan H. P., Pickering M. J. & Tanaka M. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, **118**, 172–189.

Bresnan J. 2007a. A few lessons from typology. *Linguistic Typology*, **11**(1), 297–306.

Bresnan J. 2007b. Is syntactic knowledge probabilistic ? Experiments with the English dative alternation. In S. Featherston & W. Sternefeld, Eds., *Roots : Linguistics in Search of Its Evidential Base*, p. 77–96. Berlin : Mouton de Gruyter.

Bresnan J., Cueni A., Nikitina T. & Baayen. R. H. 2007. Predicting the dative alternation. In G. Boume, I. Kraemer & J. Zwarts, Eds., *Cognitive Foundations of Interpretation*. Amsterdam : Royal Netherlands Academy of Science.

Bresnan J., Dingare S. & Manning C. D. 2001. Soft constraints mirror hard constraints : Voice and person in English and Lummi. In M. Butt & T. H. King, Eds., *Proceedings of the LFG01 Conference*, Hong Kong.

Bresnan J. & Ford M. 2010. Predicting syntax : Processing dative constructions in American and Australian varieties of English. *Language*, **86**(1), 186–213.

Bresnan J. & Hay J. 2008. Gradient grammar : An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua*, **118**(2), 245–259.

Bresnan J. & Nikitina T. 2009. The gradience of the dative alternation. In L. Uyechi & L. H. Wee, Eds., *Reality Exploration and Discovery : Pattern Interaction in Language and Life*, p. 161–184. Stanford : CSLI Publications.

Buridant C. 2000. *Grammaire nouvelle de l'ancien français*. Sedes.

Bybee J. 2006. From usage to grammar : The mind's response to repetition. *Language*, **82**(4), 711–733.

Bybee J. 2009. Language universals and usage-based theory. In M. Christiansen, C. Collins & S. Edelman, Eds., *Language Universals*, p. 17–40. Oxford : Oxford University Press.

Bybee J. 2010. *Language, Usage and Cognition*. Cambridge : Cambridge University Press.

Carletta J. 1996. Assessing agreement on classification tasks : The kappa statistic. *Computational Linguistics*, **22**(2), 249–254.

Chomsky N. 1965. *Aspects of the Theory of Syntax*. Cambridge : MIT Press.

Chomsky N. 1975. *The Logical Structure of Linguistic Theory*. Cambridge : MIT Press.

Chrupała G., Dinu G. & van Genabith J. 2008. Learning morphology with Morfette. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis & D. Tapias, Eds., *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Maroc : European Language Resources Association (ELRA).

Clark H. H. 1994. Managing problems in speaking. *Speech Communication*, **15**, 243–250.

Clark H. H. 1996. *Using Language*. Cambridge : Cambridge University Press.

Clark H. H. & Wasow T. 1998. Repeating words in spontaneous speech. *Cognitive Psychology*, **37**, 201–242.

Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Collins P. 1995. The indirect object construction in English : An informational approach. *Linguistics*, **33**(1), 35–50.

Cooper W. E. & Ross J. R. 1975. World order. In R. E. Grossman, L. J. San & T. J. Vance, Eds., *Papers from the Parasession on Functionalism*, p. 63–111. Chicago : Chicago Linguistic Society.

Cowart W. 1997. *Experimental Syntax : Applying Objective Methods to Sentence Judgments*. Thousand Oaks : Sage Publications.

Croft W. 2001. *Radical Construction Grammar*. Oxford University Press.

Davidse K. 1996. Ditransitivity and possession. In R. Hasan, C. Aoran & D. Butt, Eds., *Functional Description : Theory in Practice*, p. 85–144. Amsterdam : John Benjamins.

Delomier D. 1980. La place de l'adjectif en français, bilan des points de vue et théories du XXe siècle. *Cahiers de lexicologie*, **37**, 5–34.

Dister A. & Simon A.-C. 2008. La transcription synchronisée des corpus oraux : un aller-retour entre théorie, méthodologie et traitement informatisé. *Arena Romanistica*, **1**, 54–79.

Donohue C. & Sag I. A. 1999. Domains in Warlpiri. In *Proceedings of the 6th International Conference on Head-driven Phrase Structure Grammar (HPSG 1999)*, p. 101–106, Edinburgh.

Dubois J. & Dubois-Charlier F. 1997. *Les verbes français*. Paris : Larousse-Bordas.

Eisenberg P. 2004. *Grundriss der deutschen Grammatik 2 : Der Satz*. Stuttgart : Metzler J.B. Verlag.

Erdmann P. 1988. On the principle of 'weight' in English. In C. Duncan-Rose & T. Vennemann, Eds., *On Language, Rhetorica Phonologica Syntactica*, p. 325–339. London : Routledge.

Ertel S. 1977. Where do the subjects of sentences come from ? In S. Rosenberg, Ed., *Sentence Production. Developments in Research and Theory*. New York : Wiley.

Estival D. 1985. Syntactic priming of the passive in English. *Text*, **5**, 7–21.

Evert S. 2006. How random is a corpus ? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, **54**(2), 177–190.

Falk Y. N. 1983. Constituency, word order, and phrase structure rules. *Linguistic Analysis*, **11**, 331–360.

Fenk-Oczlon G. 1989. Word frequency and word order in freezes. *Linguistics*, **27**, 517–556.

Fodor J. D. 2002. Psycholinguistics cannot escape prosody. In *Proceedings of the Speech Prosody 2002 Conference*, Aix-en-Provence.

Ford M. 1983. A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior*, **22**(2), 203–218.

Forsgren M. 1978. *La place de l'adjectif épithète en français contemporain, étude quantitative et sémantique*. Stockholm : Almqvist & Wilksell.

Fox G. & Thuilier J. 2010. Predicting the position of attributive adjectives in the French. In *Proceedings of the 15th Student Session of ESSLLI*, p. 173–183, Copenhague.

Garretson G. 2004. Coding practices used in the project optimal typology of determiner phrases. http ://npcorpus.bu.edu/html/documentation.

Gazdar G. & Pullum G. K. 1981. Subcategorization, constituent order, and the notion of 'head'. In Moortgat, van der Hulst & Hoekstra, Eds., *The Scope of Lexical Rules*, p. 107–223. Dordrecht : Foris.

Gelman A. & Hill J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge : Cambridge University Press.

Georgia G. 1974. *Semantics and Syntactic Regularity*. Bloomington : Indiana University Press.

Gibson E. 2000. The dependency locality theory : A distance-based theory of linguistic complexity. In *Image, Language, Brain : Papers from the First Mind Articulation Project Symposium*, p. 95–126. Cambridge, MA : MIT Press.

Giry-Schneider J. 1987. *Les prédicats nominaux en français*. Genève : Librairie Droz.

Giry-Schneider J. 1996. La notion de modifieur obligatoire dans des phrases à verbe support *avoir* complexes. *Langages*, **121**, 19–34.

Glatigny M. 1967. La place des adjectifs épithètes dans deux oeuvres de Nerval. *Le français moderne*, **35**(1).

Godfrey J., Holliman E. & McDaniel. J. 1992. SWITCHBOARD : Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*, p. 517–520, San Francisco.

Goes J. 1999. *L'adjectif entre nom et verbe*. Bruxelles : De Boeck & Larcier.

Goldberg A. 1995. *Constructions : A construction Grammar Approach to Argument Structure*. Chicago : University of Chicago Press.

Goldberg A. 2006. *Constructions at Work : The Nature of Generalization in Language*. Oxford University Press.

Green G. M. 1980. Some wherefores of English inversions. *Language*, **56**(3), 582–601.

Grevisse M. & Goosse A. 2007. *Le bon usage*. 14ème édition : De Boeck Université.

Gries S. T. 2003a. Collostructions : Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, **8**(2), 209–243.

Gries S. T. 2003b. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, **1**, 1–27.

Gries S. T. 2005. Syntactic priming : A corpus-based approach. *Journal of Psycholinguistic Research*, **34**(4), 365–399.

Gries S. T. 2009. *Statistics for Linguistics with R : A Practical Introduction*. Trends in Linguistics. Studies and Monographs. Mouton de Gruyter.

Gropen J., Pinker S., Hollander M., Goldberg R. & Wilson R. 1989. The learnability and acquisition of the dative alternation. *Language*, **65**, 203–257.

Gross G. 1996. *Les expressions figées en français. Noms composés et autres locutions*. Paris : Ophrys.

Gross M. 1967. Sur une règle de « cacophonie ». *Langages*, **7**, 105–119.

Gundel J. K. 1988. Universals of topic-comment structure. In M. Hammond, E. A. Moravcsik & J. R. Wirth, Eds., *Studies in Syntactic Typology* : John Benjamins.

Guyon A. 1993. *Les adjectifs relationnels arguments de noms prédicatifs*. PhD thesis, Université Paris 7.

Harrell F. E. 2001. *Regression Modeling Strategies : With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics. Springer.

Hawkins J. 1990. A parsing theory of word order universals. *Linguistic Inquiry*, **21**, 223–261.

Hawkins J. 1994. *A Performance Theory of Order and Constituency*. Cambridge : Cambridge University Press.

Hawkins J. A. 2000. The relative order of preposition phrases in English : Going beyond manner - place - time. *Language Variation and Change*, **11**, 231–266.

Howell D. C. 1998. *Méthodes statistiques en sciences humaines*. Série Internationale. Paris : De Boeck.

Jäger G. & Rosenbach A. 2006. The winner takes it all - almost. Cumulativity in grammatical variation. *Linguistics*, **44**(5), 937–971.

Jelinek E. & Demers R. A. 1983. The agent hierarchy and voice in some Coast Salish languages. *International Journal of American Linguistics*, **49**(2), 167–185.

Jelinek E. & Demers R. A. 1994. Predicates and pronominal arguments in Straits Salish. *Language*, **70**(4), 697–736.

Judd C., McClelland G., Ryan C., Muller D. & Yzerbyt V. 2010. *Analyse des données : une approche par comparaison de modèles*. Série Internationale. Bruxelles : De Boeck.

Kamp H. 1975. Two theories about adjectives. In E. Keenan, Ed., *Formal Semantics of Natural Language*, p. 123–155. Cambridge : Cambridge University Press.

Kathol A. 2000. *Linear Syntax*. Oxford University Press.

Keenan E. L. & Comrie B. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, **8**(1), p. 63–99.

Keller F. 2000. *Gradience in Grammar : Experimental and Computational Aspects of Degrees of Grammaticality*. PhD thesis, University of Edinburgh.

Keller F. 2006. Linear optimality theory as a model of gradience in grammar. In G. Fanselow, C. Féry, R. Vogel & M. Schlesewsky, Eds., *Gradience in Grammar : Generative Perspectives*, p. 270–287. Oxford University Press.

Kempen G. & Harbusch K. 2004. A corpus study into word order variation in German subordinate clauses : Animacy affects linearization independently of grammatical function assignment. In T. Pechmann & C. Habel, Eds., *Multidisciplinary Approaches to Language Production*, p. 173–181. Berlin : Mouton de Gruyter.

Krifka M. 2004. Semantic and pragmatic conditions for the dative alternation. *Korean Journal of English Language and Linguistics*, **4**, 1–32.

Levin B. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. Cambridge, Massachusetts : MIT Press.

Levin B. & Rappaport Hovav M. 2002. What alternates in the dative alternation ? Paper presented at the 2002 Conference on Role and Reference Grammar, availiable at `http://ling.ucsd.edu/courses/lign270/Levin_Hovav_2002.pdf`.

Levin B. & Rappaport Hovav M. 2005. *Argument Realization*. Cambridge : Cambridge University Press.

Levy R. & Andrew G. 2006. Tregex and Tsurgeon : Tools for querying and manipulating tree data structures. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Gênes : European Language Resources Association (ELRA).

Lødrup H. 2007. A new account of simple and complex reflexives in Norwegian. *Journal of Comparative Germanic Linguistics*, **10**(3), 183–201.

Mallet G. 2008. *La liaison en français : descriptions et analyses dans le corpus PFC*. PhD thesis, Université Paris Ouest.

Manning C. 2003. Probabilistic syntax. In R. Bod, J. Hay & S. Jannedy, Eds., *Probabilistic Linguistics*, p. 289–341. Cambridge : MIT Press.

Manning C. D. & Schütze H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge : MIT Press.

McDonald J., Bock K. J. & Kelly M. 1993. Word and world order : Semantic, phonological and metrical determinant of serial position. *Cognitive Psychology*, **25**, 188–230.

L. McNally & C. Kennedy, Eds. 2008. *Adjectives and Adverbs : Syntax, Semantics and Discourse*. Oxford University Press.

Ménager D. 1979. *Ronsard*. Travaux d'humanisme et Renaissance. Droz.

Miller P. 1992. *Clitics and Constituents in Phrase Structure Grammar*. New York : Garland.

Miller P., Pullum G. K. & Zwicky A. M. 1997. The principle of phonology-free syntax : Four apparent counterexamples in French. *Journal of Linguistics*, **33**, 67–90.

Miller P. & Sag I. A. 1997. French clitic movement without clitics or movement. *Natural Language and Linguistic Theory*, **15**, 573–639.

Mondada L. 2000. Les effets théoriques des pratiques de transcription. *LINX*, **42**, 131–150.

Morin Y.-C. 1992. Un cas méconnu de la déclinaison de l'adjectif en français : les formes de liaison de l'adjectif antéposé. In *Mot, les mots, les bons mots. Hommage à Igor Mel'čuk*, p. 233–250. Montréal : Presses de l'Université de Montréal.

Morin Y.-C. 2003. Remarks on prenominal liaison consonants in French. In S. Ploch, Ed., *Living on the Edge. 28 Papers in Honour of Jonathan Kaye*, p. 385–400. Berlin : Mouton de Gruyter.

Morin Y.-C. 2005. La liaison relève-t-elle d'une tendance à éviter les hiatus ? Réflexions sur son évolution historique. *Langages*, **158**, 8–23.

Morolong M. & Hyman L. M. 1977. Animacy, objects and clitics in Sesotho. *Studies in African Linguistics*, **8**(3), 199–218.

Namer F. 2002. Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2002)*, p. 235–244, Nancy.

New B. 2006. Lexique 3 : une nouvelle base de données lexicales. In P. Mertens, C. Fairon, A. Dister & P. Watrin, Eds., *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, p. 892–900, Louvain : Presses Universitaires de Louvain.

New B., Pallier C. & and L. F. 2001. Une base de données lexicales du français contemporain sur internet : LEXIQUE. *L'année psychologique*, **101**, 447–462. http ://www.lexique.org.

Newmeyer F. J. 1983. *Grammatical Theory : Its Limits and Its Possibilities*. Chicago : The University of Chicago Press.

Noailly M. 1990. *Le substantif épithète*. Paris : Presse Universitaire de France.

Noailly M. 1999. *L'adjectif en français*. Paris : Ophrys.

Noailly M. à paraître. Les classes d'adjectifs. In A. Abeillé, D. Godard & A. Delaveau, Eds., *La grande grammaire du français*. Paris : Bayard.

Oehrle R. T. 1976. *The Grammar of English Dative Alternation*. PhD thesis, MIT Department of Linguistics, Cambridge, Massachusetts.

Pensado C. 1995. El complemento directo preposicional. Estado de la cuestión y bibliografia comentada. In C. Pensado, Ed., *El complemento directo preposicional*, p. 11–60. Madrid : Visor.

Pinheiro J. C. & Bates D. M. 2000. *Mixed-Effects Models in S and S-Plus*. Springer.

Pinker S. 1989. *Learnability and Cognition. The Acquisition of Argument Structure*. Cambridge, Massachusetts : MIT Press.

Polinsky M. 1996. The double object construction in spoken Eastern Armenian. *Linguistic Studies in the Non-Slavic Languages of the Commonwealth of Independent States and the Baltic Republics*, p. 307–335.

Pollard C. & Sag I. A. 1994. *Head-driven Phrase Structure Grammar*. Chicago : University of Chicago Press.

Prat-Sala M. & Branigan H. P. 2000. Discourse constraints on syntactic processing in language production : A cross-linguistic study in English and Spanish. *Journal of Memory and Language*, **42**, 168–182.

Prince A. & Smolensky P. 2004. *Optimality Theory : Constraint Interaction in Generative Grammar*. Oxford : Blackwell.

Prince E. F. 1981. Toward a taxonomy of given-new information. In P. Cole, Ed., *Radical Pragmatics*, p. 223–256. New York : Academic Press.

Pullum G. K. & Scholz B. C. 2001. On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In P. de Groote, G. Morrill & C. Retoré, Eds., *Proceedings of the 4th International Conference on Logical Aspects of Computational Linguistics*, p. 17–43, Berlin/Heidelberg : Springer-Verlag.

R Development Core Team 2011. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienne, Autriche.

Ransom E. 1979. Definiteness and animacy constraints on passive and double-object constructions in English. *Glossa*, **13**, 215–240.

Reape M. 1994. Domain union and word order variation in German. In J. Nerbonne, K. Netter & C. J. Pollard, Eds., *German in Head-Driven Phrase Structure Grammar*, p. 151–197. Standford University : CSLI Publications.

Reape M. 1996. Getting things in order. In H. Bunt & A. van Horck, Eds., *Discontinuous Constituency*, p. 209–253. Berlin, New York : Mouton de Gruyter.

Reiner E. 1968. *La place de l'adjectif épithète en français : théories traditionnelles et essai de solution*. Vienne et Stuttgart : W. Braumüller.

Rickford J., Wasow T., Mendoza-Denton N. & Espinoza J. 1995. Syntactic variation and change in progress : Loss of the verbal coda in topic-restricting *as far as* constructions. *Language*, **71**, 101–131.

Riemer N. 2009. Grammaticality as evidence and prediction in a Galilean linguistics. *Language Sciences*, **31**(5), 612–633.

Rosenbach A. 2005. Animacy versus weight as determinants of grammatical variation in English. *Language*, **81**(3), 613–644.

Sabio F. 2006. L'antéposition des compléments dans le français contemporain : l'exemple des objets directs. *Lingvisticae Investigationes*, **29**(1), 173–182.

Sabio F. 2007. La description de l' « ordre des mots » confrontée à la pluralité des usages : quelques observations sur la place des compléments en français. *Cahiers AFLS*, **13**(1), 18–32.

Sagot B. 2010. The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'10)* : European Language Resources Association (ELRA).

Schmitt C. 1987a. À propos de l'impact de la sémantique sur la séquence des compléments d'objets en français moderne. *Travaux de linguistique et de littérature*, **25**(1), 283–298.

Schmitt C. 1987b. Sémantique et prédétermination de l'ordre des mots en français contemporain. *Travaux de linguistique*, **14/15**, 21–31.

Schütze C. 1996. *The Empirical Base of Linguistics : Grammaticality Judgements and Linguistics Methodology*. Chicago : University of Chicago Press.

Seddah D., Candito M., Crabbé B. & Henestroza Anguiano E. 2012. Ubiquitous usage of a French large corpus : Processing the Est-Republicain corpus. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis, Eds., *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul : European Language Resources Association (ELRA).

Seddah D., Chrupała G., van Genabith J. & Candito M. 2010. Lemmatization and statistical lexicalized parsing of morphologically-rich languages. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2010)*, p. 85–93, Los Angeles : Association for Computational Linguistics.

Siewierska A. 1993. On the interplay of factors in the determination of word order. In J. Jacobs, A. von Stechow, W. Sternefeld & T. Vennemann, Eds., *Syntax : An International Handbook of Contemporary Research*, p. 826–846. Berlin : Mouton de Gruyter.

Snyder K. M. 2003. *The Relation between Form and Function in Distransitive Constructions*. PhD thesis, University of Pennsylvania, Philadelphia.

Sorace A. & Keller F. 2005. Gradience in linguistic data. *Lingua*, **115**(11), 1497–1524.

Stallings L. M., MacDonald M. C. & O'Seaghdha P. G. 1998. Phrasal ordering constraints in sentence production : Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, **39**(3), 392–417.

Steriade D. 1999. Lexical conservatism in French adjectival liaison. In J.-M. Authier, B. E. Bullock & L. A. Reed, Eds., *Formal Perspectives in Romance Linguistics*, p. 243–270. Amsterdam : John Benjamins.

Stubbs A. 2011. MAE and MAI : Lightweight annotation and adjudication tools. In *Proceedings of the 5th Linguistic Annotation Workshop*, p. 129–133, Portland : Association of Computational Linguistics.

Szmrecsanyi B. 2005. Language users as creatures of habit : A corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, **1**(1), 113–150.

Tanaka M., Branigan H. & Pickering M. 2011. Conceptual influences on word order and voice in sentence production : Evidence from Japanese. *Journal of Memory and Language*, **65**(3), 168–182.

Taylor J. R. 1996. *Possessives in English*. Oxford : Clarendon Press.

Thompson S. A. 1990. Information flow and dative shift in English discourse. In J. A. Edmondson, C. Feagin & P. Mühlhäusler, Eds., *Development and Diversity, Language Variation Across Space and Time*, p. 239–253, Dallas, Texas : Summer Institute of Linguistics and University of Arlington.

Thuilier J., Abeillé A. & Crabbé B. 2011. Do animate argument come first ? In *Proceedings of the conference Architectures and Mechanisms for Language Processing (AMLaP 2011)*, Paris.

Thuilier J., Fox G. & Crabbé B. 2010a. Approche quantitative en syntaxe : l'exemple de l'alternance de position de l'adjectif épithète en français. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal.

Thuilier J., Fox G. & Crabbé B. 2010b. Fréquence, longueur et préférences lexicales dans le choix de la position de l'adjectif épithète en français. In F. Neveu, V. M. Toke, T. Klingler, J. Durand, L. Mondada & S. Prévost, Eds., *Actes du 2ème Congrés Mondial de Linguistique Française 2010 (CMLF 2010)*, Nouvelle-Orléans.

Thuilier J., Fox G. & Crabbé B. 2012. Prédire la position des adjectifs épithètes en français. *Lingvisticae Investigationes*, **35**(1), 28–75.

Tily H., Gahl S., Arnon I., Snider N., Kothari A. & Bresnan J. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, **1**(2), 147–165.

Tily H., Hemforth B., Arnon I., Shuval N., Snider N. & Wasow T. 2008. Eye movements reflect comprehenders' knowledge of syntactic structure probability. In *Proceedings of the conference Architectures and Mechanisms for Language Processing (AMLaP 2008)*.

Torrego E. 1999. El complemento directo preposicional. In I. Bosque & M. Ilescu, Eds., *Gramática descriptiva de la lengua española Vol.2*, p. 1779–1805. Madrid : Espasa Calpe.

Tran M. & Maurel D. 2006. Prolexbase : un dictionnaire relationnel multilingue de noms propres. *Traitement automatique des langues*, **47**(3), 115–139.

Tranel B. 2000. Aspects de la phonologie du français et la théorie de l'optimalité. *Langue française*, **126**, 39–72.

Tsunoda T. 1985. Remarks on transitivity. *Journal of Linguistics*, **21**(2), 385–396.

Vasishth S. 2003. Quantifying processing difficulty in human sentence parsing : The role of decay, activation, and similarity-based interference. In *Proceedings of the European Cognitive Science Conference*.

Vasishth S. & Broe M. 2011. *The Foundations of Statistics : A Simulation-based Approach*. Heidelberg : Springer.

Venables W. N. & Ripley B. D. 1999. *Modern Applied Statistics with S-PLUS*. New York : Springer-Verlag.

von Heusinger K. & Kaiser G. A. 2011. Affectedness and differential object marking in Spanish. *Morphology*, **21**(1), 593–617.

Wasow T. 1997. Remarks on grammatical weight. *Language Variation and Change*, **9**, 81–105.

Wasow T. 2002. *Postverbal Behavior*. CSLI publications.

Wasow T. 2009. Gradient data and gradient grammars. In *Proceedings of the 43rd Annual Meeting of Chicago Linguistics Society*, p. 255–271.

Wasow T. & Arnold J. 2003. Post-verbal constituent ordering in English. In G. Rohdenburg & B. Mondorf, Eds., *Determinants of Grammatical Variation in English*, p. 119–154. Mouton.

Waugh L. R. 1977. *A semantic Analysis of Word Order : Position of the Adjective in French*. Leiden : E. J. Brill.

Williams R. S. 1994. A statistical analysis of English double object alternation. *Issues in Applied Linguistics*, **5**(1), 37–58.

Wilmet M. 1980. Antéposition et postposition de l'épithète qualificative en français contemporain : matériaux. *Travaux de linguistique*, **7**, 179–201.

Wilmet M. 1981. La place de l'épithète qualificative en français contemporain : étude grammaticale et stylistique. *Revue de linguistique romane*, **45**, 17–73.

Wissman M., Toutenburg H. & Shalabh 2007. *Role of Categorical Variables in Multicollinearity in the Linear Regression Model*. Rapport interne, Institut für Statistik, Ludwig-Maximilians-Universität München.

Yamamoto M. 1999. *Animacy and Reference : A Cognitive Approach to Corpus Linguistics*. John Benjamins.

Yi E., Koenig J.-P. & Mauner G. 2012. Structural repetition in sentence production conditioned by verb semantic similarity. Paper presented at the 25th annual CUNY conference on Human Sentence Processing.

Zaenen A., Carletta J., Garretson G., Bresnan J., Koontz-Garboden A., Nikitina T., O'Connor M. C. & Wasow T. 2004. Animacy encoding in English : Why and how. In B. Webber & D. K. Byron, Eds., *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, p. 118–125, Barcelona : Association for Computational Linguistics.

Zaharlick A. 1982. Tanoan studies : Passive sentences in Picuris. *The Ohio State University Working Papers in Linguistics*, **26**, 34–48.

Zipf G. K. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, USA : Harvard University Press.